


3 1761 10374386 0



Digitized by the Internet Archive
in 2023 with funding from
University of Toronto

<https://archive.org/details/31761103743860>

12
-001



34

SURVEY METHODOLOGY



Catalogue 12-001

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

JUNE 1995

•

VOLUME 21

•

NUMBER 1



Statistics
Canada

Statistique
Canada

Canada



SURVEY METHODOLOGY

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

JUNE 1995 • VOLUME 21 • NUMBER 1

Published by authority of the Minister
responsible for Statistics Canada

© Minister of Industry, 1995

All rights reserved. No part of this publication may be reproduced,
stored in a retrieval system or transmitted in any form or by any
means, electronic, mechanical, photocopying, recording or otherwise
without prior written permission from Licence Services,
Marketing Division, Statistics Canada,
Ottawa, Ontario, Canada K1A 0T6.

June 1995

Price: Canada: \$45.00
United States: US\$50.00
Other Countries: US\$55.00

Catalogue No. 12-001

ISSN 0714-0045

Ottawa



Statistics Canada
Statistique Canada

Canada

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Survey Methodology is abstracted in The Survey Statistician and Statistical Theory and Methods Abstracts and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods.

MANAGEMENT BOARD

Chairman G.J. Brackstone

Members D. Binder
B.N. Chinnappa
G.J.C. Hole
F. Mayda (Production Manager)
C. Patrick
R. Platek (Past Chairman)
D. Roy
M.P. Singh

EDITORIAL BOARD

Editor M.P. Singh, *Statistics Canada*

Associate Editors

D.R. Bellhouse, <i>University of Western Ontario</i>	D. Pfeffermann, <i>Hebrew University</i>
D. Binder, <i>Statistics Canada</i>	J.N.K. Rao, <i>Carleton University</i>
M.J. Colledge, <i>Australian Bureau of Statistics</i>	L.-P. Rivest, <i>Université Laval</i>
J.-C. Deville, <i>INSEE</i>	I. Sande, <i>Bell Communications Research, U.S.A.</i>
J.D. Drew, <i>Statistics Canada</i>	C.-E. Särndal, <i>Université de Montréal</i>
J.-J. Droesbeke, <i>Université Libre de Bruxelles</i>	W.L. Schaible, <i>U.S. Bureau of Labor Statistics</i>
W.A. Fuller, <i>Iowa State University</i>	F.J. Scheuren, <i>George Washington University</i>
M. Gonzalez, <i>U.S. Office of Management and Budget</i>	J. Sedransk, <i>State University of New York</i>
R.M. Groves, <i>University of Maryland</i>	P.J. Waite, <i>U.S. Bureau of the Census</i>
D. Holt, <i>University of Southampton</i>	J. Waksberg, <i>Westat, Inc.</i>
G. Kalton, <i>Westat, Inc.</i>	K.M. Wolter, <i>National Opinion Research Center</i>
A. Mason, <i>East-West Center</i>	A. Zaslavsky, <i>Harvard University</i>

Assistant Editors N. Laniel, M. Latouche, L. Mach, H. Mantel and D. Stukel, *Statistics Canada*

EDITORIAL POLICY

Survey Methodology publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

Survey Methodology is published twice a year. Authors are invited to submit their manuscripts in either English or French to the Editor, Dr. M.P. Singh, Household Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6. Four nonreturnable copies of each manuscript prepared following the guidelines given in the Journal are requested.

Subscription Rates

The price of Survey Methodology (Catalogue No. 12-001) is \$45 per year in Canada, US \$50 in the United States, and US \$55 per year for other countries. Subscription order should be sent to Publication Sales, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6. A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, and the Statistical Society of Canada.

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Volume 21, Number 1, June 1995

CONTENTS

In This Issue	1
A Tribute to Stanley L. Warner	
Introductory Remarks by C.-E. SÄRNDAL	3
Principal Publications and Papers of Stanley L. Warner	5
S.E. FIENBERG and N. JAZAIRI	
Stanley Warner's Contributions to Statistically Balanced Information Technology	7
D.R. BELLHOUSE	
Estimation of Correlation in Randomized Response	13
N.S. MANGAT, R.SINGH, S. SINGH, D.R. BELLHOUSE and H.B. KASHANI	
On Efficiency of Using Distinct Respondents in a Randomized Response Survey.....	21
<hr/>	
P. LAVALLÉE	
Cross-sectional Weighting of Longitudinal Surveys of Individuals and Households Using the Weight Share Method	25
G. KALTON and J.M. BRICK	
Weighting Schemes for Household Panel Surveys	33
P. DICK	
Modelling Net Undercoverage in the 1991 Canadian Census	45
J.J. KIM, A. ZASLAVSKY and R. BLODGETT	
Between-State Heterogeneity of Undercount Rates and Surrogate Variables in the 1990 U.S. Census	55
F.J. BREIDT	
Markov Chain Designs for One-Per-Stratum Sampling	63
G. MEEDEN	
Median Estimation Using Auxiliary Information	71
B. HULLIGER	
Outlier Robust Horvitz-Thompson Estimators	79
R. IACHAN and S.S. KEMP	
Visitor Sample Surveys	89

In This Issue

This issue of the *Survey Methodology* journal contains a special memorial section in honour of Stanley L. Warner, which includes an introduction by C.-E. Särndal, a bibliography of Warner's principal publications and papers, organized by topic, and three papers dealing with areas in which Warner was a pioneer. The first paper, by Fienberg and Jazairi, summarizes Warner's work in the area of statistically balanced information technology, in which the goal is to develop statistical procedures to ensure that different positions regarding a policy or an issue are fairly and adequately represented in a debate or decision. The other two papers, by Bellhouse, and by Mangat, *et al.*, are in the area of randomized response. The paper by Bellhouse begins with an overview of Warner's contributions to randomized response and then discusses the problem of estimation of a correlation coefficient using data from a randomized response survey. Three randomized response setups are considered: unrelated question, additive constant, and multiplicative constant. The paper by Mangat *et al.* compares the efficiencies of with and without replacement sampling in the context of randomized response. The papers by Fienberg and Jazairi and by Bellhouse are based on presentations given at a special session in memory of Warner at the meetings of the Statistical Society of Canada in Banff in 1994.

The next two papers, by Lavallée and by Kalton and Brick, discuss weighting schemes for cross-sectional estimation in panel surveys.

Lavallée presents the Weight Share Method, used in cross-sectional estimation for longitudinal surveys, in a more general context. He demonstrates the unbiasedness of the method and obtains a general expression for the variance of the estimator of a total. Then the author illustrates the method by applying it in the context of the Survey of Labour and Income Dynamics of Statistics Canada. The estimation of variance is also discussed.

Kalton and Brick describe weighting schemes for cross-sectional analysis of later waves of a household panel survey using data for all households for whom data are collected. These weighting schemes can accommodate new entrants to the population who move in to live with members of the original population, but not other new entrants. The authors discuss cases where the schemes are optimal as well as further weighting adjustments to compensate for nonresponse and noncoverage.

Small area techniques for estimation of net undercoverage of persons in population censuses are discussed by Dick in the Canadian context and by Kim, Zaslavsky and Blodgett in the U.S. context.

The paper by Dick describes modelling that was done in order to produce estimates of census net undercoverage of persons within age-sex-province categories for the 1991 Canadian Census using data from the Reverse Record Check and the Overcoverage Study. An Empirical Bayes model for direct estimates of adjustment factors is formulated and used to obtain smoothed estimates of those adjustment factors. The smoothed estimates of net missed persons are then raked to match the direct estimates of national age-sex group and provincial totals, which are considered to be of good quality.

Kim, Zaslavsky and Blodgett describe the two analyses performed to test the "synthetic assumption" of homogeneity of undercount rates between parts of different states falling in the same poststratum for the 1990 U.S. Census. In the first analysis, the distributions of five "surrogate variables" that, like undercount, were related to the census-taking process, were investigated using a large extract from the census. In the second analysis, the distribution of undercount was analyzed using the Post Enumeration Survey data.

Breidt presents Markov chain designs for one-per-stratum sampling which includes systematic sampling, stratified simple random sampling and balanced systematic sampling as special cases. He introduces new designs that are shown to be competitive, in terms of the efficiency of the Horvitz-Thompson estimator of a total, with standard one-per-stratum designs under a variety of superpopulation models. Theoretical and numerical comparisons are provided.

Meeden considers the problem of estimation of the median when an auxiliary variable is available. He uses a non-informative Bayesian approach based on a Polya posterior for the ratios of the variable of interest to the covariate. The resulting estimator is empirically compared to a number of alternatives in terms of bias and average absolute error for a variety of real and synthetic populations. The Polya posterior is also to be used to generate interval estimates which are evaluated empirically. Robustness of the procedure to moderate departures from the assumptions is also considered.

Hulliger develops design-based M-estimators for samples with unequal inclusion probabilities. He expresses the Horvitz-Thompson (HT) estimator as a least square functional and then makes it robust against outliers through M-estimators, analogous to the robustification of least square estimators in linear models for infinite populations. He also provides an approximation to the sampling variance of this robustified HT-estimator and its estimate. The results of the Monte-Carlo study confirm that the robustified HT-estimators outperform the HT-estimator in many outlier situations.

Iachan and Kemp describe the sampling designs for two visitor sample surveys of recreational users of parks, a survey of National Park Service area users over a one year period, and a survey of users of three river basin in the Pittsburgh area. The potential problems associated with sampling in both time and space are described, and the ways in which the designs of these two surveys meet these challenges are compared and contrasted.

The Editor

Stanley L. Warner

1928-1992

Born and educated in the United States, Stanley Warner received a Ph.D. in Economics from Northwestern University in 1961. In 1971, he moved to Canada, where he was to pursue the rest of his academic career as professor at York University in the Department of Economics and the Faculty of Administrative Studies. He died suddenly in August 1992 at the age of 63.

Stan was a highly original thinker. His statistical research was guided by relevance and common sense. In his memory, a session was organized at the Statistical Society of Canada meetings at Banff in 1994. Two of the papers which follow were presented at that occasion, namely, David Bellhouse: "Estimation of Correlation in Randomized Response" and Stephen E. Fienberg and Nuri Jazairi: "Stanley Warner's Contributions to Statistically Balanced Information Technology." They deal with two areas where Stan made pathbreaking contributions.

Stan Warner is known to many, especially to survey statisticians, as the man who invented randomized response. This technique is used in surveys with sensitive questions as in a survey concerning drug usage among high school students. The objective is to eliminate two embarrassing nonsampling errors: measurement error and nonresponse. Bias caused by such errors is a problem that has haunted statisticians since the beginnings of survey taking. Standard methods exist to "adjust for" such bias. They may reduce but do not eliminate the bias. Stan, disregarding the conventional wisdom, tackled the problem in a completely unorthodox way and came up with an unbiased solution useful at least for some surveys. His seminal article, "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias", appeared in the *Journal of the American Statistical Association* in 1965.

Randomized response, if carried out according to the intentions, guarantees the anonymity of the respondent: a "yes" answer will not identify the respondent since his or her question is selected at random. Unbiased estimation of the "yes" proportion in the population is nevertheless possible (usually at the price of some increase in variance) because the survey taker knows the number of "yes" responses as well as the probability with which the random choice device selects the sensitive question (rather than its opposite or a completely unrelated question).

The idea struck the imagination of many statisticians. Numerous modifications, refinements and extensions were given, as evidenced through the 1988 bibliography of contributions to randomized response put together by Chaudhuri and Mukerjee. Why this stream of papers?

That nonresponse bias was always an important practical problem without a satisfactory solution is not the whole explanation. There is also the fact that it seemed like magic, even to experienced statisticians, that valid answers could be obtained without knowing what question had been asked of the respondent. Of course now that the idea exists, it is not hard to explain; in fact it has become a favorite classroom example, useful even in an elementary statistics course, to show students the powers of statistical reasoning.

Implementing randomized response in practice requires some special arrangements, including a choice device that randomly selects a question. As time went by, Stan realized that the technique had to be adapted to modern low cost data gathering. As late as 1989 at the International Statistical Institute meetings in Paris, he presented a "quick randomized response" version suitable for telephone surveys and touch tone entry, thereby reducing time and cost.

The paper by David Bellhouse traces the development of randomized response in more detail.

Later, Stan's interest focused on statistical procedures for balanced information, which occupied him from about 1975 and on. In the last few years of his life he was working on a book on the topic; the manuscript is now being prepared for publication.

The goal of balanced information technology is to give statistical procedures to ensure that different positions regarding a policy or an issue can be fairly and adequately represented. Stan's first paper on this topic, entitled "Advocate Scoring for Unbiased Information", appeared in the *Journal of the American Statistical Association* in 1975. It deals with the situation in which advocates are each to give *pro* and *con* information regarding an issue to a number of individuals who, after exposure to the information, are to express opinions for or against the issue. Each advocate is charged with using the given data to prepare separate *pro* and *con* cases.

Decisions are frequently made on information provided by advocates; this occurs in government, education, law, etc. One can imagine that Stan was concerned with the incomplete and sometimes arbitrary way in which quantitative information is used in decision making of the utmost importance, including political summits, where prestige, mistrust and political consideration would often prevent the parties from meeting on even ground.

The paper by Stephen E. Fienberg and Nuri Jazairi presents the main ideas of this work, which is probably less familiar to statisticians.

Another example of Stan's creative spirit is the fact that he developed, with his wife, a musician, a system for music notation which is in wide use.

I did not know Stan Warner at the time of his original work on randomized response but keep vivid memories of conversations with him later in his career. These occasions were not that numerous; however, each left a strong impression on me. The warmth, modesty and unassuming

manner of this highly original man could not fail to make an impression. Stan was a real scholar, not a run-of-the-mill researcher: he did not hesitate to follow the sometimes lonely route laid out to him by belief in ideas that were truly his own.

C.-E. Särndal

Principal Publications and Papers of Stanley L. Warner

1928-1992

1. PH.D. THESIS AND RELATED PUBLICATIONS

- (1962) *Stochastic Choice of Mode in Urban Travel. A Study in Binary Choice*, Evanston, Illinois: Northwestern University Press.
- (1963) Multivariate regression of dummy variates under normality assumptions. *Journal of the American Statistical Association*, 58, 1054-1063.
- (1967) Asymptotic variances for dummy variate regression under normality assumptions. *Journal of the American Statistical Association*, 62, 1305-1314.

2. RANDOMIZED RESPONSE

- (1965) Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60, 63-69.
- (1971) The linear randomized response model. *Journal of the American Statistical Association*, 66, 884-888.
- (1976) Optimal randomized response models. *International Statistical Review*, 44, 205-212.
- (1976) With F. Leysieffer. Respondent jeopardy and optimal randomized response models. *Journal of the American Statistical Association*, 71, 649-656.
- (1979) Extended randomized response applications. In *Ethical and Legal Problems in Applied Social Research*, (Eds. R. Boruch, J. Ross and J.C. Cecil), Evanston, Illinois: Northwestern University Press.
- (1986) The omitted digit randomized-response model for telephone and other applications. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 441-443.
- (1989) Using randomized response for forecasting dimensions of the AIDS problem. Invited keynote presentation. *The Ninth International Symposium on Forecasting*, Vancouver.
- (1989) Quick randomized response. *Proceedings of the 47th Session, International Statistical Institute*, Paris, Contributed paper, 431-432.

3. BALANCED INFORMATION

- (1975) Advocate scoring for unbiased information. *Journal of the American Statistical Association*, 70, 15-22.
 - (1977) Advocate scoring design for technological and social policy assessment. *Proceedings of the 41st Session, International Statistical Institute*, New Delhi, Book 3 – Invited Papers, 373-379.
 - (1979) Subjective information in statistics. *Proceedings of the Business and Economics Section, American Statistical Association*, 558-563.
 - (1981) Balanced information, the Pickering Airport experiment. *The Review of Economics and Statistics*, LXII, 256-262.
 - (1984) The overlapping information survey model for evaluating summary information. *Proceedings of the Social Statistics Section, American Statistical Association*, 581-584.
 - (1985) Applications of the overlapping information model. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 401-403.
 - (1985) The overlapping information model for measuring summary information. *Proceedings of the 45th Session, International Statistical Institute*, Amsterdam, Contributed paper, 49-50.
 - (1987) Identifying rational opinion-formation with the overlapping model. In *Applied Probability, Stochastic Processes and Sampling Theory*, (Eds. I.B. MacNeill and G.J. Umphrey). Dordrecht, The Netherlands: D. Reidel, 323-329.
 - (1987) Using test populations to develop balanced agenda. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 441-443.
- Statistically Balanced Information Technology*, (Manuscript of Book).

4. CLINICAL TRIALS

- (1981) Post-treatment randomization in clinical trials. *Proceedings, Statistics Eighty One Canada*, Concordia University Canada.

- (1981) Post-treatment randomization in clinical research. *Proceedings of the Social Statistics Section, American Statistical Association*, 233-236.
- (1982) Post-treatment randomization extensions for medical and educational research. *Proceedings of the Social Statistics Section, American Statistical Association*, 410-413.
- (1983) Post-treatment randomization estimates. *Proceedings of the 44th Session, International Statistical Institute*, Madrid, 464-468.

5. OTHERS

- (1958) With R.L. Andreano. Professor Bain and barriers to new competition. *Journal of Industrial Economics*, 66-78.
- (1965) Cost models, errors in variables and economies of scale in trucking. *The Cost of Trucking, Econometric Analysis*, Dubuque: William C. Brown and Co., 1-46.
- (1983) With W.D. Cook and L. Seiford. Preference ranking models: Conditions for equivalence. *Journal of Mathematical Sociology*, 9, 125-137.

Stanley Warner's Contributions to Statistically Balanced Information Technology

STEPHEN E. FIENBERG and NURI JAZAIRI¹

ABSTRACT

Stanley Warner was widely known for the creation of the randomized response technique for asking sensitive questions in surveys. Over almost two decades he also formulated and developed statistical methodology for another problem, that of deriving balanced information in advocacy settings so that both positions regarding a policy issue can be fairly and adequately represented. We review this work, including two survey applications implemented by Warner in which he applied the methodology, and we set the ideas into the context of current methodological thinking.

KEY WORDS: Advocate scoring; Bayes' Theorem; Embedded experiment; Logistic regression; Survey analysis.

1. INTRODUCTION

Consider some recent controversial public or professional issues such as:

1. Should Canada endorse the North American Free Trade Agreement?
2. Should Quebec secede from the Canadian Federation?
3. Should the American Statistical Association adopt a program to certify statisticians?
4. Should smoking be banned in all restaurants in Ottawa?

The discussions and debates surrounding such issues often reflect highly polarized positions and "pro" and "con" arguments can strongly influence the opinions of individuals in the relevant populations of interest (e.g., Canadian residents, ASA members, those who frequent restaurants in Ottawa). How to think about the presentation of such advocacy information in a balanced fashion is the topic of this paper.

It has often been said that only a small fraction of scientists make a truly novel research contribution once in their lifetime. Far fewer are responsible for multiple innovations. Stanley Warner is well-known for his creation and development of the randomized response model for surveys and that contribution has been widely hailed as a major development in statistics. What is less well known is his truly novel approach to the problem of balanced information in advocacy settings, on which he worked over a period of almost 20 years. As York University colleagues of Warner's at the time of his death in 1992, we know how seriously he took the obligation of statistics and statisticians to deal with such complex problems, and this work is one example of how he attempted to fulfill the obligation.

Our goal in this paper is to reintroduce Warner's ideas on the topic of balanced information in advocacy settings

to the profession and to demonstrate how they fit into current survey practice and methodological thinking. In Section 2, we present his basic approach to the advocacy problem and we describe the statistical model he chose to focus upon (Warner 1975). In Sections 3 and 4, we discuss embellishments of the basic approach which he presented in subsequent papers (e.g., see Warner 1981, 1984, 1985, 1987a), and we end by describing how Warner continued to pursue this research program up until the time of his death. In the process, we also stress the importance Warner attached to the application of his ideas.

2. THE BASIC PROBLEM

In a thoughtful, well-argued, yet provocative 1975 paper in the *Journal of the American Statistical Association*, Stanley Warner first presented the issue of measuring the impact of advocacy and balance on public opinion in connection with controversial issues. He did so by asking (and then answering) a pair of interrelated questions:

1. How can we estimate what the population would conclude on issue were each of the members provided with balanced information on the topic?

Warner's idea for answering this question was to use advocates to present summaries of arguments, both "pro" and "con," and to implement this in a factorial experimental design to different samples, and in the process achieve information about a balanced presentation. This then leads rather naturally to the second question:

2. How can we rate or score advocates in such settings?

He developed his formulation to answer the two questions simultaneously and, in doing so, he used *both* economical and statistical arguments. In this paper, we focus on the statistical portion of his arguments and refer the interested reader to Warner's paper for the economic details.

¹ Stephen E. Fienberg, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213, U.S.A.; Nuri Jazairi, Department of Economics, York University, North York, Ontario, Canada, M3J 1P3.

Consider a pair of advocates or advocate teams whose role it is to brief individuals on the arguments associated with a controversial issue, H . Let $P(H)$ and $P(\bar{H})$ denote the proportion of the number of subjects in a given population “for” and “against” issue H . Let F_i and A_j denote “pro” and “con” presentations of advocates i and j , respectively for $i, j = 1, 2$. Let $P(H | F_i, A_j)$ and $P(\bar{H} | F_i, A_j)$ denote the number of subjects “for” and “against” issue H after hearing “pro” case from advocate i and “con” case from advocate j .

Warner defined the “net information” associated with F_i and A_j as

$$I(F_i, A_j) = \ln[P(H | F_i, A_j)/P(\bar{H} | F_i, A_j)] - \ln[P(H)/P(\bar{H})]. \quad (1)$$

Formula (1) is, of course, the logarithm of the Bayes factor, or what Good (1950) called the *weight of evidence*. While Warner recognized the evocative nature of the use of Bayes’ Theorem here, his approach towards its use was purely frequentist.

Similarly, Warner defined the net information associated with “pro” and “con” cases of F_i and A_j , separately:

$$I(F_i) = \ln[P(H | F_i)/P(\bar{H} | F_i)] - \ln[P(H)/P(\bar{H})], \quad (2)$$

$$I(A_j) = \ln[P(H | A_j)/P(\bar{H} | A_j)] - \ln[P(H)/P(\bar{H})]. \quad (3)$$

The simplest assumption we can make relating the joint and marginal information quantities is that of independence of the “pro” and “con” cases,

$$I(F_i, A_j) = I(F_i) + I(A_j), \quad (4)$$

for $i = 1, 2$, and $j = 1, 2$. This assumption allows for some direct comparisons and, as we shall see, can be checked empirically.

In order to ensure that the advocates fairly treat both “pro” and “con” positions, Warner proposed to reward them on the basis of the sum of the net information they provided, *i.e.*

$$I(F_i) + I(A_j). \quad (5)$$

Economic theory, Warner argued, suggests that rewarding advocates in this fashion will lead them to at least strive to approximate the “unbiased information” associated with maximization under resource constraints. Thus we need to estimate the quantity in expression (5) along with the posterior odds implied by unbiased information:

$$P(H | F', A') / P(\bar{H} | F', A'). \quad (6)$$

“Balance” in design for data collection was the key to Warner’s plan for estimation.

Warner’s estimation plan was linked to his application. The controversial issue was the completion of the north-south Spadina Expressway in Toronto (Warner’s home city). The original first section of the expressway was constructed in 1966 and, after much debate, the remainder of the project was canceled in 1971. Two years later, in 1973, Warner conducted a survey to learn what proportions of the population of registered voters of Metropolitan Toronto were for or against the original expressway plan. He took a random sample of 1,360 registered voters (1% of the corresponding population) divided into 8 equal subsamples of size 170. Two advocate teams prepared written positions, both “pro” and “con” the expressway, and one of each was included in the mailing. The order of presentation of the two written positions was also varied producing a $2 \times 2 \times 2$ experimental design with the first variable corresponding to who prepared the “pro” brief, the second to who prepared the “con” brief, and the third to the order of presentation (“pro” first or “con” first). Advocates were paid a basic fee and a larger amount was set aside to be paid to the team with the “best combined score.” This is an excellent example of a factorial experiment embedded within a survey, and fits well with the spirit of embedding described in Fienberg and Tanur (1988).

Table 1

Sample Preferences for Spadina Expressway After Information by Advocates

Sample	i	j	k	For	Against	Undecided	Total	p_{ijk}	n_{ijk}
1	1	1	1	22	4	1	27	.846	26
2	1	1	2	18	9	2	29	.666	27
3	1	2	1	26	8	0	34	.764	34
4	1	2	2	21	11	1	33	.656	32
5	2	1	1	28	10	1	39	.736	38
6	2	1	2	14	11	1	26	.560	25
7	2	2	1	19	16	1	36	.542	35
8	2	2	2	19	17	2	38	.527	36

Source: Warner (1975).

In the cover letter, Warner asked respondents to return prepaid postcards indicating their preferences after reviewing the briefs. At the cut off date, 262 cards had been returned for a response rate of about 20%. The resulting data, in Table 1, are reproduced from Warner (1975).

Let p_{ijk} be the true proportion of the population “for” the expressway, in group (i, j, k) . Then, with an additive term for order of presentation, the model of expression (1) becomes

$$\ln[p_{ijk}/(1 - p_{ijk})] = \ln[P(H)/P(\bar{H})] + I(F_i) + I(A_j) + D_k. \quad (7)$$

We now recognize expression (7) as a linear logit model, and the sampling scheme as product-binomial (ignoring the correction for the 0.2% sampling fraction). Of course when Warner did this work it preceded the existence of a monograph by Bishop, Fienberg, and Holland (1975), and was virtually concurrent with Nelder and Wedderburn's (1972) paper on generalized linear models. Thus his paper made no reference to the now extensive literature on logit and loglinear models.

To estimate the parameters in expression (7), Warner used weighted least squares, which yields both estimated coefficients and standard errors. Instead of dealing directly with the parameters in expression (7), he redefined them, in part to simplify computation and in part to aid in their interpretation:

$$\beta_1 = \ln \frac{P(H)}{P(\bar{H})} + \frac{I(F_1) - I(A_1)}{2} + \frac{I(F_2) - I(A_2)}{2} + \frac{D_1 + D_2}{2}, \quad (8)$$

$$\beta_2 = I(F_1) - I(F_2), \quad (9)$$

$$\beta_3 = I(A_1) - I(A_2), \quad (10)$$

$$\beta_4 = D_1 - D_2. \quad (11)$$

The coefficient β_1 is an "intercept" or normalizing parameter, while β_2 , β_3 , and $\beta_2 + \beta_3$ measure the performance of the advocate teams, and β_4 measures the order effect. The net information provided by team 1 is $\beta_1 + .5(\beta_2 - \beta_3)$, and that provided by team 2 is $\beta_1 - .5(\beta_2 - \beta_3)$. The difference in net influence is thus $\beta_2 - \beta_3$.

Table 2

Weighted Least Squares Estimates of Theoretical Parameters

Parameter	Estimate	Approx. Std. Error
β_1	.712	.139
β_2	.648	.277
β_3	-.383	.275
β_4	.528	.274
$\beta_2 + \beta_3$.264	.386
$\beta_1 + .5\beta_2 - .5\beta_3$	1.228	.266
$\beta_1 - .5\beta_2 + .5\beta_3$.196	.215
$\beta_2 - \beta_3$	1.032	.395

Source: Warner (1975).

We reproduce Warner's estimation results in Table 2. We have double-checked the estimated values in Table 2 using the generalized model routines in *S+*, which utilize a version of iteratively weighted least squares (maximum

likelihood in this case). Our logit model computations agree with Warner's to two decimal places. The residual deviance for this model equals 1.95 with 4 d.f., indicative of a remarkably good fit and offering strong support for the reasonableness of the independence assumption of expression (4).

In interpreting the results in Table 2, Warner noted that his economic analysis leads to the conclusion that the overall proportion of the population in favour of *H* when presented with unbiased information lies between the "pure" estimates for the 2 advocate teams, or in the present instances (.55, .77). These bounds correspond to the estimates in the 2nd and 3rd last lines of the table. As was clear from Table 1, no matter how we combine "pro" and "con" arguments, the majority in each subgroup favored completion of the expressway. Warner observed that we might be tempted to use $\hat{\beta}_1$ to produce a "best estimate" of the value of *p* corresponding to unbiased information, but he argued for a higher value, since Team 1 is superior to Team 2 in terms of total information, *i.e.*, $\hat{\beta}_2 - \hat{\beta}_3 > 0$. (The superiority of Team 1 is quite evident from a quick examination of Table 1 and does not require the full analysis.)

Warner ended his 1975 paper by pointing out all of the shortcomings of his small experiment, and his initial modelling efforts. What we can observe in retrospect is the way in which he was able to attack a very complicated public policy and survey problem using a simple but ingenuous model, as well as a rigorous estimation scheme built on the solid framework of a factorial experiment embedded in a sample survey, and then actually applying the methodology to produce an answer for a real problem.

It is worth noting that the first version of this paper was submitted for publication to *JASA* in June 1972, before Warner had actually carried out the empirical study on the Spadina Expressway controversy. Over two years passed before he resubmitted a revised version of the paper with the detailed example. Even well-known authors with innovative ideas often struggle to have their work published in major statistical journals, and a compelling empirical application is always of help.

3. EXTENSIONS AND A SECOND APPLICATION

Warner extended his balanced information approach in a second paper (Warner 1981), focusing on yet another application. This paper also signals a substantial change in Warner's thinking about statistics and probability, towards a subjective Bayesian approach and away from the classical approaches that he stressed in his early career. While the reported analyses are still frequentist in nature, Warner used, at least informally, the assessments of prior probabilities in a manner that fits rather naturally with the Bayesian formulation of expression (1) above.

In March 1972, the Canadian Federal Government announced a plan to build a second Toronto International Airport to the east of the city in Pickering, Ontario. This led to considerable controversy. In 1974, the government appointed a 3-person commission of inquiry. Warner carried out a concurrent but independent survey experiment. The question he posed was whether or not the Pickering Airport should be built before the year 2000. The general structure of the experiment was similar to the previous one on the Spadina Expressway controversy, but with some differences:

- (i) This time his study population was economists.
- (ii) He incorporated 2 “neutral” control sub-samples, which received neither “pro” nor “con” statements.
- (iii) Respondents in the 8 experimental subsamples gave probability assessments (instead of 0-1 values) after assessing the advocacy positions. Those in the control groups also gave their probability assessment.

The test population was limited to those economists who belonged to the Canadian Economic Association or who could be identified as professors or lecturers in an economics department in a Canadian university. The survey was done via mail in two stages – the first identified those willing to read detailed briefs and “report opinions regarding an undisclosed federal project,” and the second mailing divided those willing to participate into 10 sub-samples, corresponding to the $2 \times 2 \times 2$ design of Section 2 plus the 2 control samples consistent of those who were asked for their opinions without briefs. A total of 726 economists participated in the experiment. In Table 3, we provide Warner's summary of the data for the 8 experimental subsamples in which he aggregated the posterior judgments into three groups according to whether they were substantially greater, nearly equal, or substantially less than 0.5. The data have been further aggregated across the 8 experiment groups. The results have been post-stratified according to whether the economists were professors, graduate students, or others. The data on “prior beliefs” come from a combination of the two control groups.

Table 3
Test Population Opinions on Pickering Airport

	Professors		Students		Others		Totals	
	Before Briefs	After Briefs	Before Briefs	After Briefs	Before Briefs	After Briefs	Before Briefs	After Briefs
For	9 (.143)	58 (.266)	9 (.257)	32 (.288)	11 (.180)	71 (.298)	29 (.182)	161 (.284)
Against	32 (.508)	155 (.711)	12 (.343)	72 (.648)	36 (.590)	160 (.672)	80 (.503)	387 (.683)
Undecided	22 (.349)	5 (.023)	14 (.400)	7 (.063)	14 (.230)	7 (.029)	50 (.315)	19 (.033)
Totals	63 (1.000)	218 (1.000)	35 (1.000)	111 (1.000)	61 (1.000)	238 (1.000)	159 (1.000)	567 (1.000)

Source: Warner (1981).

Note that all three groups had substantial negative opinions about the proposed airport, a posteriori, and that the differences in proportions of undecided between the experimental and control subsamples provide evidence that the advocacy briefs affected public opinion on the issue. Warner's formal statistical analysis of the data focused solely on the 8 experimental subsamples and utilized three variants of the formal model in expression (9) and the reparameterization of expressions (10) through (13):

- (i) A logit structure similar to that in Warner (1975) based on the aggregation in Table 3, with “undecideds” in effect *imputed* as belonging in either the “pro” or “con” categories with probability 0.5. He called this a *Simple Aggregate Influence* model.
- (ii) A more direct approach, which averaged the posterior assessments to get “aggregate proportions” in favor, and then treated these observed proportions as if they were binomial. He called this a *Weighted Aggregate Influence* model.
- (iii) A two-stage model, which first used individual-level assessments breaking up the range of 0 to 1 into 17 levels, and then a “variable coefficient” regression model analysis. He referred to this as an *Average Disaggregate Influence* model.

Each analysis involved the use of a different form of weighted least squares to estimate the coefficients of interest.

In Table 4, we provide Warner's estimated coefficients under all three models and analyses. The results are similar across models and we can summarize the findings as follows:

- (a) Team 2 clearly presented the strongest case ($\hat{\beta}_2$ and $\hat{\beta}_3$ are both positive and similar for all three columns).
- (b) The estimated aggregate influence for Team 2 is $[\hat{\beta}_1 + .5(\hat{\beta}_2 - \hat{\beta}_3)] = -0.688$ corresponding to an estimated proportion in favor of the airport project of $\hat{p} = 0.355$.
- (c) The disaggregate influence for Team 2 corresponds to an estimated proportion in favor of the airport project of $\hat{p} = 0.355$.
- (d) The effect of order of presentation ($\hat{\beta}_4$) suggests that the brief appearing first in the enclosures had greater impact, and is consistent with the hypothesis that the “previous information favoring one position serves to discount new information against that position.”

It turns out that the advisor for Team 1 felt that construction of the airport could not be defended and this seriously handicapped the “pro” efforts of Team 1 (something reflected in the estimates of β_2).

Table 4

Estimated Case Influence for Pickering Airport Experiment

Parameters	Simple Aggregate Influence	Weighted Aggregate Influence	Average Disaggregate Influence
β_1	-.857 (.093)	-.529 (.047)	-.736 (.065)
β_2	.485 (.188)	.337 (.097)	.462 (.132)
β_3	.147 (.188)	.146 (.097)	.187 (.132)
β_4	.313 (.186)	.209 (.095)	.307 (.129)
N	8	8	567

Source: Warner (1981).

4. OVERLAPPING INFORMATION

Warner worried that the information used in the Pickering Airport survey experiment involved an overlap between the “pro” and “con” cases, and there was also an overlap between the prior information available to the respondents and that presented by the advocates. He turned to this question several years later in Warner (1984, 1985), using a formal argument drawn from sampling theory.

Warner’s idea was to consider N pieces of independent information being used to influence the proposition in question. Let Y_{ij} be the information content seen by the i -th person in the j -th piece of information. Then the prior odds for the i -th individual is

$$\ln[p_i/(1 - p_i)] = \sum_{j \in A(i)} Y_{ij}, \quad (12)$$

where $A(i)$ is the collection of information seen by the i -th person prior to the presentation of the advocacy arguments. If $A(i)$ is empty, the initial log-odds for individual i should be 0 and thus $p_i = 0.5$.

The presented “pro” and “con” summaries draw on a subset, S , of m out of the N units. Suppose that participants act “rationally” and are not further influenced by data which has been seen before. The added information is then

$$\sum_{j \in S} Y_{ij} - \sum_{j \in A(i) \cap S} Y_{ij}. \quad (13)$$

If the m units of information are randomly selected without replacement from the total N , then this implies that we can treat the units in $A(i) \cap S$ as having been selected at random without replacement. We can treat the information from these overlapping units as following a hypergeometric distribution, and then we rewrite expression (1) as

$$I_i = m\bar{S}_i - m/N[p_i/(1 - p_i)] + \epsilon, \quad (14)$$

where I_i is the net information in the summary for the i -th individual, \bar{S}_i is the average of the Y_{ij} ’s for those data units $j \in S$, and the error term ϵ has zero conditional expectation, *i.e.*,

$$E\{\epsilon \mid [p_i/(1 - p_i)]\} = 0. \quad (15)$$

If we group subjects in an advocacy experiment exposed to the same “pro” and “con” briefs according to the values of p_i , then differences in net information should be related to $[p_i/(1 - p_i)]$ according to equation (14). Warner (1984) did this with additional simplifying assumptions and then analyzed the data from the Pickering experiment using assessments, the control groups and the Team 2 “pro” and “con” group, aggregated according to whether the respondent was a professor, a student, or other. The problem with using the data from the Pickering experiment is that we are in effect matching the individuals in the control group and the experimental group. What we really want is both the prior and the posterior assessments from the same individual (Warner 1987b).

Warner (1985, 1987a) returned to this theme of overlapping information, and he extended the model of expression (14) to take the form:

$$I_i = m\bar{S}_i + D_i r_i (Z_i - U_i), \quad (16)$$

where we have in effect replaced the coefficient $-m/N$ in expression (14) by $D_i r_i$, where $D_i \geq -1$ is a discount factor and $E(r_i) = m/N$. He then showed how to estimate the coefficients in this “random coefficients” regression model using generalized least squares, under various assumptions about the correlations among the quantities in (16).

He then applied the approach to a new set of data collected for a telephone survey of Carleton University students on the question of whether an elected Canadian Senate would be preferable to the existing appointed Senate. The interviewees were asked for their opinion on the issue expressed as a probability. They were then presented with a 6-sentence summary of a television debate on the topic, and asked to reconsider their probability assessments. Of the 417 participants, 316 gave prior probabilities different from 0 and 1. Of this group, 163 actually changed their assessment, and overall the average log-odds after the summary was virtually the same as it was before, but with a slightly smaller variance. Warner actually fit the model to the data, and the fitted equations were consistent with the notion of partial discounting of the information that they had already seen.

This was essentially Warner’s last published contribution to the topic of balanced information assessment. At the time of his death, Warner was hard at work on a book-length

manuscript whose title, *Statistically Balanced Information Technology*, suggests that he was attempting to synthesize and extend his ideas on the topic. Unfortunately, we have only been able to locate the early chapters of this book and these include just the introductory ideas on probability and regression that he expected to utilize in the later chapters.

5. FURTHER OBSERVATIONS

Warner's balanced information technology addresses the common problem of adversarial policy advocacy which may give rise to confusion and incorrect decisions because of imbalance in the presentation of the relevant facts. Examples of how the adversarial approach to dispute resolution in a legal setting could have distorting effects on questions of scientific fact are discussed in Fienberg (1989; see especially Appendix H by Vidmar). Among the responses to this situation have been repeated proposals to establish a science court to ensure balance in organizing the information relevant to a factual dispute and reaching decisions. In these proposals, the science court itself is an adversarial system, but based on well-defined procedures for the selection of issues, advocates, and judges designed to ensure impartiality and minimize the effects of personal bias. Warner's approach outlined here is a formal way to achieve precisely this kind of impartial result.

Warner's progression through the various stages of the work on balanced information was paralleled by a shift in his outlook on the foundations of statistics. He was trained as an economist and a classical statistician and his early statistical contributions, including the work on randomized response models, were all set in a frequentist statistical framework. The 1975 paper on advocate scoring represented his first step towards a subjectivist perspective and, with each successive paper, he added further elements of the Bayesian approach. In Warner (1979), Stan articulated this shift in thinking and it is especially apparent in the early chapters of his unpublished book. At the May 1992 annual meeting of the Statistical Society of Canada in Edmonton, his last public lecture, Stan described devices for the solicitation of probabilities that he had been developing for the book.

We can only speculate about how Stan's subjectivist synthesis of balanced information technology would have looked had he been able to complete the book. But given the depth of his commitment to the Bayesian approach and its recent methodological innovations, we expect that it would have included a hierarchical generalized linear model approach and utilized the latest developments in Markov Chain Monte Carlo simulation techniques.

Stanley Warner was constantly using the ideas from his research in the classroom and in reflecting back upon the work described here, he noted:

"... almost all of the basic elements of an elementary statistics course are to some degree represented in these procedures, and the problems in modeling and design that are suggested could be considered at quite an advanced level", (Warner 1987b).

The statistics profession has lost a true innovator and a great educator. We count ourselves amongst Stanley Warner's students and we continue to learn from his work.

ACKNOWLEDGMENTS

The preparation of this paper was supported in part by a grant from the Natural Sciences and Engineering Research Council of Canada to the first author at York University.

REFERENCES

- BISHOP, Y.M.M., FIENBERG, S.E., and HOLLAND, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge: MIT Press.
- FIENBERG, S.E., and TANUR, J.M. (1988). From the inside out and outside in: Combining experimental and sampling structures. *Canadian Journal of Statistics*, 16, 135-151.
- FIENBERG, S.E., (Ed.) (1989). *The Evolving Role of Statistical Assessments as Evidence in the Courts*. New York: Springer-Verlag.
- GOOD, I.J. (1950). *Probability and the Weighing of Evidence*. London: Griffin.
- NELDER, J.A., and WEDDERBURN, R.W.M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, A*, 135, 370-384.
- WARNER, S.L. (1975). Advocate scoring for unbiased information. *Journal of the American Statistical Association*, 70, 15-22.
- WARNER, S.L. (1979). Subjective information in statistics. *Proceedings of the Business and Economics Section, American Statistical Association*, 558-563.
- WARNER, S.L. (1981). Balanced information, The Pickering Airport experiment. *The Review of Economics and Statistics*, LXII, 256-262.
- WARNER, S.L. (1984). An overlapping information survey model for evaluating summary information. *Proceedings of the Social Statistics Section, American Statistical Association*, 581-584.
- WARNER, S.L. (1985). Applications of the overlapping-information model. *Proceedings of the Section on Survey Research Methods, American Statistical Society*, 401-403.
- WARNER, S.L. (1987a). Identifying rational opinion-formation with the overlapping information model. In *Applied Probability, Stochastic Processes, and Sampling Theory*. (Eds. I.B. MacNeill and G.J. Umphrey). Dordrecht, The Netherlands: Reidel, 323-329.
- WARNER, S.L. (1987b). Using test populations to develop balanced agenda. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 441-443.
- WARNER, S.L. (1992). *Statistically Balanced Information Technology*. Unpublished manuscript.

Estimation of Correlation in Randomized Response

D.R. BELLHOUSE¹

ABSTRACT

Stanley Warner's contributions to randomized response are reviewed. Following this review, a linear model, based on random permutation models, is developed to include many known randomized response designs as special cases. Under this model optimal estimators for finite population variances and covariances are obtained within a general class of quadratic design-unbiased estimators. From these results an estimator of the finite population correlation is obtained. Three randomized response designs are examined in particular: (i) the unrelated questions model of Greenberg *et al.* (1969); (ii) the additive constants model of Pollock and Bek (1976); and (iii) the multiplicative constants model of Pollock and Bek (1976). Simple models for response bias are presented to illustrate the effect of this bias on estimation of the correlation.

KEY WORDS: Additive constants model; Linear models; Multiplicative constants model; Response bias; Unrelated question model; Variance estimation.

1. A BRIEF OVERVIEW OF WARNER'S CONTRIBUTIONS TO RANDOMIZED RESPONSE

Randomized response is a technique used to elicit responses to sensitive questions. It was developed thirty years ago by Stanley Warner (Warner 1965) to estimate a proportion under a simple random sampling design with replacement. The development was a substantial intellectual achievement requiring much originality of thought. How does one get truthful responses to sensitive questions? Warner's solution was to get the response without the interviewer knowing whether the sensitive question had actually been asked. He devised the probabilistic structure to the questioning so that an estimate of the required proportion could be obtained. In Warner's original formulation the population is divided into two mutually exclusive and exhaustive groups, A and B. It is of interest to estimate the proportion π of the population belonging to group A. To do this, a spinner is constructed with a face marked with the letters A and B. The construction is such that the spinner points to the letter A with probability p and to B with probability $1 - p$. The interviewee spins the spinner and is required only to say yes or no according to whether or not the spinner points to the interviewee's correct membership group. The with replacement design allows estimation of π by maximum likelihood.

This very original idea has received substantial attention over the past thirty years. Since Warner's original work, several randomized response techniques have been suggested for the estimation of a proportion or set of proportions as in polytomous data, or for the estimation of a population mean with continuous data. A variation on Warner's original theme is asking the sensitive question or an

unrelated question with probabilities p and $1 - p$ respectively. This was originally due to Greenberg *et al.* (1969). Other variations with continuous data include adding a random variable to the response to the sensitive question or multiplying the response by a random variable. The underlying theme to any of these techniques is the masking of the original response in such a way that the sensitive information cannot be attributed to any single respondent but that information on the sensitive attribute can be extracted from the whole sample. A substantial literature, including a monograph by Chaudhuri and Mukerjee (1988), has grown up around these techniques. Nathan (1988) has provided a fairly comprehensive bibliography of this literature. Umesh and Peterson (1991) have given several detailed examples from very diverse areas of the application and applicability of the techniques of randomized response.

With several different randomized response techniques, the question arises as to how to compare the different methods. Minimization of variance cannot be the sole criterion. Each method is designed to protect the privacy of the respondent. A gain in efficiency, in terms of variance, by the choice of different values of the probabilities in the randomizing device, or by the choice of one randomized response method over another, could lead to jeopardizing the privacy of the respondents. In response to this, Leysieffer and Warner (1976) and Warner (1976) formulated natural measures of respondent jeopardy. These measures are related to the probability of the interviewer being able to infer the interviewee's response to the sensitive attribute. The theory of respondent jeopardy is reviewed in Chaudhuri and Mukerjee (1988) and some practical considerations regarding respondent jeopardy are reviewed in Umesh and Peterson (1991).

¹ D.R. Bellhouse, Department of Statistical and Actuarial Sciences, University of Western Ontario, London, Ontario, N6A 5B7.

Stanley Warner made two other contributions to the literature of randomized response. The first contribution is directly related to the results obtained here. With the explosion of new ideas and new techniques in randomized response, Warner (1971) formulated a linear model which unified the theory. Most of the randomized response techniques at that time could be put in his linear model framework. The second contribution was in response to the growing use of telephone interviewing. Stem and Steinhorst (1984) described randomized response methods applicable to telephone interviewing and to mail questionnaires. Warner (1986) suggested practical natural randomizing devices, such as the serial numbers on paper money, for use in telephone interviewing.

The major topics in randomized response methodology are: the development of randomized response techniques, the comparison of these techniques through the concept of respondent jeopardy, the construction of reasonable randomizing devices, the development of a unified theory of randomized response, and the validation of randomized response techniques through field studies. Stanley Warner's contributions to randomized response touch on most of these major developments in the subject. Moreover, most of these contributions were substantial and influential. He is the originator of the technique. His original setup of a dichotomous population was quickly generalized to a polytomous one and to populations with continuous measurement. New randomized response techniques continue to be developed. Warner was at the forefront of evaluating randomized response designs through the modeling of respondent jeopardy. His work in the development of a unified linear model for randomized response designs was the foundation on which a unified theory of randomized response has been built.

2. INTRODUCTION TO ESTIMATION OF CORRELATION

Consider a finite population of size N with two measurements of interest x_j and y_j for $j = 1, \dots, N$. It is of interest to estimate the finite population correlation

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y},$$

where $\sigma_{xy} = \sum (x_j - \bar{X})(y_j - \bar{Y})/N$ is the finite population covariance between the variables x and y and where σ_x^2 and σ_y^2 are the finite population variances of the variables x and y respectively. To estimate ρ a sample of fixed size n is chosen with probability $P(s)$ from the finite population where s denotes the set of finite population units chosen for the sample. The expectation operator with respect to the sampling design $P(s)$ is denoted by E_p . Estimators for ρ are obtained by replacing σ_x^2 , σ_y^2 and σ_{xy}

by their respective estimators, unbiased or biased, optimal in some sense or otherwise.

To illustrate the general results obtained here for estimation of the finite population correlation coefficient, three particular randomized response techniques will be considered:

- (i) The unrelated questions model due to Greenberg *et al.* (1969). The sensitive question is asked with probability p and an unrelated question which is not sensitive is asked with probability $1 - p$. For estimation of the mean it is assumed that the finite population mean \bar{X} of the unrelated question is known. For estimation of variance it is also assumed that σ_x^2 is known.
- (ii) The additive constants model due to Pollock and Bek (1976). The outcome of a random variable from a known probability distribution is added to the value of the response to the sensitive question.
- (iii) The multiplicative constants model due to Pollock and Bek (1976). The value of the response to the sensitive question is multiplied by the outcome of a random variable from a known probability distribution.

Edgell *et al.* (1986) have provided estimators for ρ under the unrelated questions model and the additive constants model.

Most randomized response designs that have been considered have assumed that the sampling design is simple random sampling either with or without replacement. Since the results obtained here are under a fixed size design, the simple random sampling design assumed here is without replacement.

Assume that both x and y are sensitive variables. Consequently, a randomized response technique is used to obtain information on both these variables. Let w_j and z_j , for $j \in s$ be the sampled measurements that are obtained. Let u_j and v_j for $j = 1, \dots, N$ be the nonsensitive measurements associated with x_j and y_j respectively. Under the unrelated question model (randomized response model (i)) u_j and v_j are the responses to the unrelated questions for the j -th individual. Under the additive constants model or the multiplicative constants model (randomized response models (ii) or (iii)) u_j and v_j are the j -th outcomes of random variables from two, possibly different, known probability distributions.

3. RANDOM PERMUTATION MODELS

Several models for the finite population measurements have been put forward in the survey sampling literature. Here attention is focused on the random permutation models of Rao (1975) and Rao and Bellhouse (1978). One compelling reason for using these models is that the model parameters have a direct interpretation in the finite population of interest since model parameters in random

permutation models are also finite population parameters. In the simplest context for random permutation models it is assumed that the N -dimensional vector of finite population measurements is a random permutation of an N -dimensional vector of fixed numbers. Rao (1975) has shown how this assumption leads to a linear model. Bellhouse (1980) extended this model to randomized response designs under unequal probability sampling.

The model and associated designs applicable to unequal probability sampling are not easily applicable to estimation of variances and covariances either with or without a randomized response. Consequently, a special case of the model in Bellhouse (1980) is given here. In the model which follows there are two different expectation operators at work which together yield a composite expectation E_m . These expectation operators are: E_r , the expectation operator with respect to the randomizing device, and E_{rp} , the expectation operator with respect to the random permutation model. The composite expectations $E_m = E_{rp}E_r$ and $E = E_mE_p$. For the random permutation model we assume that the pairs (x_j, y_j) , $j = 1, \dots, N$ are a random permutation of a set of N fixed pairs of numbers, say (p_j, q_j) , $j = 1, \dots, N$. This is a special case of model (4.1) in Rao and Bellhouse (1978); the more general model in Rao and Bellhouse (1978) was used in double sampling and sampling on two occasions. The unrelated questions randomized response model (randomized response model (i)) requires an additional assumption that the quadruples (x_j, y_j, u_j, v_j) , $j = 1, \dots, N$ are a random permutation of a set of N fixed quadruples of numbers, say (p_j, q_j, r_j, t_j) , $j = 1, \dots, N$.

Assume that the randomizing device coupled with the random permutation model leads to the following linear model:

$$\begin{aligned} w_j &= \alpha_1 + \beta_1 \bar{X} + e_{1j} \\ z_j &= \alpha_2 + \beta_2 \bar{Y} + e_{2j}, \end{aligned} \quad (1)$$

for $j = 1, \dots, N$ where \bar{X} and \bar{Y} are the finite population means of the x and y measurements respectively and where for $j = 1, \dots, N$

$$E_m(e_{1j}) = E_m(e_{2j}) = 0,$$

$$E_m(e_{1j}^2) = \phi_1 \sigma_x^2 + \psi_{01} + \psi_{11} \bar{X} + \psi_{21} \bar{X}^2,$$

$$E_m(e_{2j}^2) = \phi_2 \sigma_y^2 + \psi_{02} + \psi_{12} \bar{Y} + \psi_{22} \bar{Y}^2,$$

$$E_m(e_{1j} e_{1k}) = \delta_1 \sigma_x^2 + \lambda_1, \quad E_m(e_{2j} e_{2k}) = \delta_2 \sigma_y^2 + \lambda_2, \\ \text{for } j \neq k,$$

$$E_m(e_{1j} e_{2j}) = \phi_3 \sigma_{xy} + \psi_3, \quad \text{and}$$

$$E_m(e_{1j} e_{2k}) = \delta_3 \sigma_{xy} + \lambda_3, \quad \text{for } j \neq k. \quad (2)$$

and all other higher moments are independent of j . In the model given by (1) and (2), the α 's, λ 's, ϕ 's, ψ 's and δ 's are all known constants. The finite populations variances and covariances of the sensitive questions, σ_x^2 , σ_y^2 and σ_{xy} are all unknown.

For the unrelated questions model (randomized response model (i)) assume that the randomizing schemes on the two sensitive questions are independent and that sensitive question i , $i = 1, 2$, is asked with probability p_i and the associated nonsensitive questions with probability $1 - p_i$. Assume further that the sensitive questions are unrelated to the nonsensitive questions so that $\sigma_{xu} = \sigma_{yv} = \sigma_{xv} = \sigma_{yu} = 0$. This assumption is unnecessary under simple random sampling with replacement. When, in addition, a random permutation model is assumed on the quadruple (x_j, y_j, u_j, v_j) then in the model given by (1) and (2):

$$\begin{aligned} \alpha_1 &= (1 - p_1) \bar{U}, \quad \beta_1 = p_1, \quad \alpha_2 = (1 - p_2) \bar{V}, \quad \beta_2 = p_2, \\ \phi_1 &= p_1, \quad \psi_{01} = (1 - p_1) \sigma_u^2 + p_1 (1 - p_1) \bar{U}^2, \\ \psi_{11} &= -2p_1 (1 - p_1) \bar{U}, \quad \psi_{21} = p_1 (1 - p_1), \\ \phi_2 &= p_2, \quad \psi_{02} = (1 - p_2) \sigma_v^2 + p_2 (1 - p_2) \bar{V}^2, \\ \psi_{12} &= -2p_2 (1 - p_2) \bar{V}, \quad \psi_{22} = p_2 (1 - p_2), \\ \delta_1 &= -p_1^2 / (N - 1), \quad \lambda_1 = -(1 - p_1)^2 \sigma_u^2 / (N - 1), \\ \delta_2 &= -p_2^2 / (N - 1), \quad \lambda_2 = -(1 - p_2)^2 \sigma_v^2 / (N - 1), \\ \phi_3 &= p_1 p_2, \quad \delta_3 = -\phi_3 / (N - 1), \\ \psi_3 &= (1 - p_1) (1 - p_2) \sigma_{uv}, \\ \text{and } \lambda_3 &= -\psi_3 / (N - 1). \quad (3) \end{aligned}$$

Note that the model assumptions require that the finite population variance-covariance matrix of the nonsensitive questions is known as well as the finite population means.

For the additive constants model (randomized response model (ii)) assume that the random variables u and v that are added to the value of the responses to the two sensitive questions are independent with means μ_u and μ_v and variances σ_u^2 and σ_v^2 respectively. When the random permutation model is assumed on the pair (x_j, y_j) then in the model given by (1) and (2):

$$\begin{aligned} \alpha_1 &= \mu_u, \quad \beta_1 = 1, \quad \alpha_2 = \mu_v, \quad \beta_2 = 1, \\ \phi_1 &= \phi_2 = \phi_3 = 1, \quad \psi_{01} = \sigma_u^2, \quad \psi_{02} = \sigma_v^2, \\ \delta_1 &= \delta_2 = \delta_3 = -1 / (N - 1), \\ \psi_{11} &= \psi_{21} = \psi_{12} = \psi_{22} = \psi_3 = \lambda_1 = \lambda_2 = \lambda_3 = 0. \end{aligned} \quad (4)$$

In the multiplicative constants model, two independent random variables, u and v with means μ_u and μ_v and variances σ_u^2 and σ_v^2 respectively, are multiplied respectively by the value of the response on the x -variable and the y -variable. When the random permutation model is assumed on the pair (x_j, y_j) then in the model given by (1) and (2):

$$\begin{aligned}\alpha_1 &= \alpha_2 = 0, \beta_1 = \mu_u, \beta_2 = \mu_v, \\ \phi_1 &= \mu_u^2 + \sigma_u^2, \phi_2 = \mu_v^2 + \sigma_v^2, \phi_3 = \mu_v \mu_u, \\ \psi_{21} &= \sigma_u^2, \psi_{22} = \sigma_v^2, \\ \delta_1 &= -\mu_u^2/(N-1), \delta_2 = -\mu_v^2/(N-1), \\ \delta_3 &= -\mu_u \mu_v/(N-1), \text{ and} \\ \psi_{01} &= \psi_{11} = \psi_{02} = \psi_{12} = \psi_3 = \lambda_1 = \lambda_2 = \lambda_3 = 0.\end{aligned}\quad (5)$$

4. ESTIMATION OF VARIANCE AND COVARIANCE

Consider estimation of σ_y^2 so that the appropriate data are z_j for units $j \in s$. The general class of quadratic estimators of σ_y^2 is of the form:

$$e_{bs} = b_{s..} + \sum_{j \in s} b_{sj.} z_j + \sum_{j \in s} b_{sjj} z_j^2 + \sum_{i \neq j \in s} b_{sij} z_i z_j, \quad (6)$$

where the coefficients of the z 's are defined for all s , all $j \in s$ and all pairs $(i, j) \in s$.

In the context of randomized response, an estimator e_b in the class defined by (6) is design-unbiased for σ_y^2 if $E_p E_r(e_b) = \sigma_y^2$ and is pm -unbiased if $E(e_b) = \sigma_y^2$. Conditions under which an estimator e_b is pm -unbiased are obtained upon taking the expectation E of (6) under (1) and (2). On equating coefficients in \bar{Y}^0 , \bar{Y}^1 , \bar{Y}^2 and σ_y^2 four equations in four unknowns are obtained. The solution to these four equations yields the following conditions under which estimators in the class defined by (6) are pm -biased for σ_y^2 :

$$E_p \left(\sum_{j \in s} b_{sjj} \right) = \frac{\beta_2^2}{\beta_2^2(\phi_2 - \delta_2) - \delta_2 \psi_{22}} = A_2, \quad (7)$$

$$\begin{aligned}E_p \left(\sum_{i \neq j \in s} b_{sij} \right) &= -\frac{\beta_2^2 + \psi_{22}}{\beta_2^2(\phi_2 - \delta_2) - \delta_2 \psi_{22}} = \\ &= -(A_2 + B_2), \quad (8)\end{aligned}$$

$$E_p \left(\sum_{j \in s} b_{sj.} \right) = \frac{(2\alpha_2 \psi_{22} - \beta_2 \psi_{12})}{\beta_2^2(\phi_2 - \delta_2) - \delta_2 \psi_{22}} = C_2, \quad (9)$$

and

$$\begin{aligned}E_p(b_{s..}) &= \frac{\lambda_2(\beta_2^2 + \psi_{22}) - (\alpha_2^2 \psi_{22} - \alpha_2 \beta_2 \psi_{12} + \beta_2^2 \psi_{02})}{\beta_2^2(\phi_2 - \delta_2) - \delta_2 \psi_{22}} \\ &= D_2. \quad (10)\end{aligned}$$

In order to obtain the optimal estimator we need to define an associated class of quadratic estimators of 0. This is given by

$$e_{cs} = c_{s..} + \sum_{j \in s} c_{sj.} z_j + \sum_{j \in s} c_{sjj} z_j^2 + \sum_{i \neq j \in s} c_{sij} z_i z_j.$$

The conditions for an estimator e_c in this class to be pm -biased for 0 are

$$\begin{aligned}E_p(c_{s..}) &= E_p \left(\sum_{j \in s} c_{sj.} \right) = E_p \left(\sum_{j \in s} c_{sjj} \right) = \\ E_p \left(\sum_{i \neq j \in s} c_{sij} \right) &= 0. \quad (11)\end{aligned}$$

Derivation of the minimum variance quadratic design-unbiased estimator of σ_y^2 follows along the same lines as that used for the finite population mean by Rao and Bellhouse (1978) for cases without randomized response and by Bellhouse (1980) for cases with randomized response. The covariance $E(e_b e_c)$ under the composite expectation is determined under the model such that only expectations of the form E_p remain to be determined. From this expression the coefficients b are set to make $E(e_b e_c) = 0$ under the conditions in (11). The values of the coefficients b are then determined from the conditions in (7) through (10). From a theorem on minimum variance unbiased estimation of Rao (1952), the resulting estimator is the optimal pm -unbiased estimator of σ_y^2 . If there exists a design such that this estimator is also design-unbiased for σ_y^2 , then by arguments similar to those given in Theorem (2.4) of Rao and Bellhouse (1978), the estimator is also the optimal design-unbiased estimator of σ_y^2 . We present results for pm -unbiased estimators first (Theorems 1 and 2) and then present results for design-unbiased estimators under the three randomized response schemes.

Theorem 1. Under the model defined by (1) and (2) and for any design of fixed size n , the pm -variance of e_b , $E_{rp}[E_p E_r(e_b - \sigma_y^2)^2] = E(e_b - \sigma_y^2)^2$, is minimized for the estimator given by

$$(A_2 + B_2)s_z^2 - B_2 \frac{1}{n} \sum_{j \in s} z_j^2 + C_2 \bar{z} + D_2, \quad (12)$$

where \bar{z} is the sample mean of the data and

$$s_z^2 = \frac{1}{n-1} \sum_{j \in s} (z_j - \bar{z})^2$$

is the sample variance of the data obtained through randomized response where A_2 , B_2 , C_2 and D_2 are defined in (7) through (10) respectively.

Proof. Under the model given by (1) and (2) the covariance $E(e_b e_c)$ is algebraically quite lengthy but may be expressed in the following form:

$$b^T G c + H, \quad (13)$$

where b^T is the vector

$$\left[E_p(b_{s..}), E_p\left(\sum_{j \in s} b_{sj.}\right), E_p\left(\sum_{j \in s} b_{sij}\right), E_p\left(\sum_{i \neq j \in s} b_{sij}\right) \right], \quad (14)$$

and c^T is the same as (14) with the b 's replaced by c 's. The 4×4 matrix G in (13) contains functions of the first order moments of z_j and the second order moments of e_{2j} in (1). The expression H in (13) is a sum of terms of the form

$$\kappa \sum b_{sij} c_{skl}, \quad (15)$$

where the summation symbol is up to a quadruple sum, where the subscripts of b could be replaced by a dot (.) and where κ is a function of second through fourth order moments of e_{2j} in (1). Note that these moments are all independent of j . In (15) the sum is a single sum over $j \in s$ when, for example, the subscripts $i = j = k = l$ or when $i = k$ and j and l are replaced by dots. The sum is a double sum over $i \neq k \in s$ when, for example, $i \neq k$ and j and l are replaced by dots. This process continues to the quadruple sum in which $i \neq j \neq k \neq l$. From (11) $E(e_b e_c)$ reduces to 0 if $b_{s..} = h_1$, $b_{sj.} = h_2$, $b_{sij} = h_3$, and $b_{sij} = h_4$, where the h_i are constants. From (7) through (10) and the fact that the design is of fixed size we obtain

$$b_{s..} = D_2, b_{sj.} = C_2/n, b_{sij} = -\frac{A_2 + B_2}{n(n-1)}, b_{sij} = A_2/n,$$

so that the estimator in (12) minimizes the variance in the pm -unbiased class of quadratic estimators of σ_y^2 . Q.E.D.

By the same arguments

$$(A_1 + B_1)s_w^2 - B_1 \frac{1}{n} \sum_{j \in s} w_j^2 + C_1 \bar{w} + D_1, \quad (16)$$

is the optimal pm -unbiased estimator for σ_x^2 where

$$A_1 = \frac{\beta_1^2}{\beta_1^2(\phi_1 - \delta_1) - \delta_1\psi_{21}},$$

$$B_1 = \frac{\beta_1^2 + \psi_{21}}{\beta_1^2(\phi_1 - \delta_1) - \delta_1\psi_{21}},$$

$$C_1 = \frac{(2\alpha_1\psi_{21} - \beta_2\psi_{11})}{\beta_1^2(\phi_1 - \delta_1) - \delta_1\psi_{21}}, \text{ and}$$

$$D_1 = \frac{\lambda_1(\beta_1^2 + \psi_{21}) - (\alpha_1^2\psi_{21} - \alpha_1\beta_1\psi_{11} + \beta_1^2\psi_{01})}{\beta_1^2(\phi_1 - \delta_1) - \delta_1\psi_{21}}.$$

The same technique can be used to estimate the covariance σ_{xy} . The general class of quadratic estimators of σ_{xy} is of the form

$$e_{ds} = d_s + \sum_{j \in s} d_{1sj} z_j + \sum_{j \in s} d_{2sj} w_j + \sum_{i \neq j \in s} d_{sij} w_i z_j,$$

where the coefficients of the w 's and z 's are defined for all s , all $j \in s$ and all pairs $(i, j) \in s$. The result on the covariance is stated without proof in

Theorem 2. Under the model defined by (1) and (2) and for any design of fixed size n , the pm -variance of e_d , $E_{rp}[E_p E_r(e_d - \sigma_{xy})^2] = E(e_d - \sigma_{xy})^2$, is minimized for the estimator given by

$$\frac{s_{wz} - (\psi_3 - \lambda_3)}{\phi_3 - \delta_3}, \quad (17)$$

where

$$s_{wz} = \frac{1}{n-1} \sum_{j \in s} (w_j - \bar{w})(z_j - \bar{z})$$

is the sample covariance between w and z .

An estimator for ρ is obtained from (12), (16) and (17). In the additive constants randomized response model (randomized response model (ii)) the estimator of ρ is given by

$$\hat{\rho}_{ac} = \frac{s_{wz}}{\sqrt{(s_w^2 - \sigma_u^2)(s_z^2 - \sigma_v^2)}}. \quad (18)$$

This is the same as the estimator obtained by Edgell *et al.* (1986). Under the multiplicative constants model (randomized response model (iii)) the estimator reduces to

$$\hat{\rho}_{mc} =$$

$$\frac{s_{wz}}{\sqrt{s_w^2 - \frac{\sigma_u^2/\mu_u^2}{1 + \sigma_u^2/\mu_u^2} \frac{1}{n} \sum_{j \in S} w_j^2} \sqrt{s_z^2 - \frac{\sigma_v^2/\mu_v^2}{1 + \sigma_v^2/\mu_v^2} \frac{1}{n} \sum_{j \in S} z_j^2}}, \quad (19)$$

for $\mu_u \neq 0$ and $\mu_v \neq 0$. When $\mu_u = 0$ the coefficient of $\sum w_j^2$ is $1/n$ and when $\mu_v \neq 0$ the coefficient of $\sum z_j^2$ is $1/n$. The estimator for ρ under the unrelated questions model (randomized response model (i)) is

$$\hat{\rho}_{uq} = \frac{s_{wz} - \frac{(1-p_1)(1-p_2)}{p_1 p_2} S_{uv}}{\sqrt{\hat{S}_x^2 \hat{S}_y^2}}, \quad (20)$$

where $S_{uv} = N\sigma_{uv}/(N-1)$ and where

$$\begin{aligned} \hat{S}_x^2 &= s_w^2 - (1-p_1) \frac{1}{n} \sum_{j \in S} w_j^2 + 2(1-p_1)\bar{U}\bar{w} - \\ &\quad (1-p_1)\bar{U}^2 - (1-p_1)\sigma_u^2 \left(p_1 + \frac{1-p_1}{N-1} \right) \end{aligned}$$

and

$$\begin{aligned} \hat{S}_y^2 &= s_z^2 - (1-p_2) \frac{1}{n} \sum_{j \in S} z_j^2 + 2(1-p_2)\bar{V}\bar{z} - \\ &\quad (1-p_2)\bar{V}^2 - (1-p_2)\sigma_v^2 \left(p_2 + \frac{1-p_2}{N-1} \right). \end{aligned}$$

When $p_1 = p_2$ this may be compared to the estimator in Edgell *et al.* (1986). The resulting estimator for $\hat{\rho}_{uq}$ differs from the estimator in Edgell *et al.* (1986) who assume that $\sigma_{uv} = 0$. They also use biased estimators of σ_x^2 and σ_y^2 . Edgell *et al.*'s estimator for σ_y^2 is obtained by writing the design variance of \bar{z} under simple random sampling with replacement as

$$\sigma_z^2/n = \sum_{j=1}^N (z_j - \bar{Z})^2/(Nn). \quad (21)$$

The design variance of \bar{z} under the randomizing device is

$$[p_2\sigma_y^2 + (1-p_2)\sigma_v^2 + p_2(1-p_2)(\bar{Y} - \bar{V})^2]/n. \quad (22)$$

Expression (22) is found in Greenberg *et al.* (1971). The estimator for σ_y^2 is found by equating (22) to the left hand side of (21), by substituting sample the estimator of σ_z^2 and the randomized response estimator of \bar{Y} in the resulting equation, and then by solving for σ_y^2 .

Each of the estimators of the finite population variances and covariance, which are the components of $\hat{\rho}$ in (18), (19) and (20), are design-unbiased under the appropriate randomized response model for any design with joint inclusion probability for units i and j given by $\pi_{ij} = n(n-1)/[N(N-1)]$. Consequently, each estimator is the optimal design-unbiased estimator for its finite population parameter counterpart. To obtain the appropriate unbiased estimators in (18), multiply the numerator and denominator each by $(N-1)/N$. The resulting numerator is design-unbiased for σ_{xy} and the expressions under the square root sign in the denominator of (18) are unbiased for σ_x^2 and σ_y^2 . In (19) it is necessary to multiply the numerator and denominator by $(N-1)/[N\mu_u\mu_v]$ in order to obtain the correct form of the design-unbiased estimators. The correct estimators are obtained in (20) when the multiplier is $(N-1)/(Np_1p_2)$.

In any of the randomized response designs, the simplest estimate of the variance of $\hat{\rho}$ is the jackknife estimate of variance. Jackknife estimates of variance for $\hat{\rho}$ can be obtained from formulae (4.2.3) or (4.2.5) in Wolter (1985).

5. EFFECT OF RESPONSE BIAS

In the additive constants model, the respondent is asked to add a random variable u to x and an independent random variable v to y . Instead, the respondent may add different independent random variables, say u' and v' . The means and variances of u' and v' may differ from those of u and v . It is reasonable to assume, however, that $\sigma_{u'}^2 \geq \sigma_u^2$ and $\sigma_{v'}^2 \geq \sigma_v^2$. One example in which this situation might occur is the following. The respondent does not want to add on the outcome of a random variable near to the mean of the distribution of the random variable. In this case the distribution of response bias could be modelled by the original distribution with an interval around the mean in which any outcome from the original distribution which falls in the chosen interval is set to one of the end points of the interval. On taking separately the expectations of the numerator and the expression under each of the square root signs in the denominator of (18) the expression

$$\frac{\sigma_{xy}}{\sqrt{\sigma_x^2 + \sigma_{u'}^2 - \sigma_u^2} \sqrt{\sigma_y^2 + \sigma_{v'}^2 - \sigma_v^2}}, \quad (23)$$

is obtained. From (23) it may be noted that the response bias leads to an estimate of correlation lower than the true value.

The multiplicative constants model is the same as the additive constants model with the exception that the responses to the sensitive questions are multiplied by the random variables. As in the response bias model for additive constants, assume that u' and v' are used by the

respondent instead of u and v . Then on taking separately the expectations of the numerator and the expressions under each of the square root signs in the denominator of (19) the expression

$$\frac{\sigma_{xy}}{\sqrt{\sigma_x^2 + \frac{\sum_{j=1}^N x_j^2}{N\mu_u^2} \frac{\sigma_u^2 \mu_{u'}^2 - \sigma_{u'}^2 \mu_u^2}{\sigma_u^2 + \mu_u^2}}} \sqrt{\sigma_y^2 + \frac{\sum_{j=1}^N y_j^2}{N\mu_v^2} \frac{\sigma_v^2 \mu_{v'}^2 - \sigma_{v'}^2 \mu_v^2}{\sigma_v^2 + \mu_v^2}}, \quad (24)$$

is obtained. If $\mu_u = \mu_{u'}$, $\mu_v = \mu_{v'}$, $\sigma_{u'}^2 \geq \sigma_u^2$ and $\sigma_{v'}^2 \geq \sigma_v^2$, as in the case of the additive constants model, then from (24) the response bias leads to an overestimate of the correlation.

In the unrelated questions model a reasonable model for response bias is to assume that the sensitive questions are answered with probability $p'_1 < p_1$ and $p'_2 < p_2$. In general the effect of this response bias is dependent on the relative values of the various probabilities, the means and variances of the sensitive questions, and the means and variances of the nonsensitive questions. Under simple random sampling without replacement and the response bias model, the design expectation of the numerator of (20) is given by

$$p'_1 p'_2 \left[S_{xy} + \frac{(1-p'_1)(1-p'_2) - (1-p_1)(1-p_2)}{p'_1 p'_2} S_{uv} \right],$$

which is greater than $p'_1 p'_2 S_{xy}$. Likewise the design expectation of S_x^2 in (20) is

$$\begin{aligned} S_x^2 [p_1'^2 + (N-1)p'_1(p_1 - p'_1)/N] \\ + (p_1 - p'_1) S_u^2 [p_1' - (p'_1 + 2p_1 - 2)/N] \\ + p'_1(p_1 - p'_1)(\bar{X} - \bar{U})^2, \end{aligned}$$

which is greater than $p_1'^2 S_x^2$ when N is large. If $S_{uv} = 0$, then the response bias leads to an underestimate of the correlation.

ACKNOWLEDGMENTS

This paper is dedicated to the memory of Stan Warner and was written for the Stanley Warner Memorial Session at the Statistical Society of Canada meetings in Banff, Alberta, May 1994. The research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

- BELLHOUSE, D.R. (1980). Linear models for randomized response designs. *Journal of the American Statistical Association*, 75, 1001-1004.
- CHAUDHURI, A., and MUKERJEE, R. (1988). *Randomized Response: Theory and Techniques*. New York: Marcel Dekker.
- EDGELL, S.E., HIMMELFARB, S., and CIRA, D.J. (1986). Statistical efficiency of using two quantitative randomized response techniques to estimate correlation. *Psychological Bulletin*, 100, 251-256.
- GREENBERG, B.G., ABUL-ELA, A.A., SIMMONS, W.R., and HORVITZ, D.G. (1969). The unrelated question randomized response model: theoretical framework. *Journal of the American Statistical Association*, 64, 520-539.
- GREENBERG, B.G., KUEBLER, R.R., ABERNATHY, J.R., and HORVITZ, D.G. (1971). Application of the randomized response technique in obtaining quantitative data. *Journal of the American Statistical Association*, 66, 243-250.
- LEYSIEFFER, F.W., and WARNER, S.L. (1976). Respondent jeopardy and optimal designs in randomized response models. *Journal of the American Statistical Association*, 71, 649-656.
- NATHAN, G. (1988). A bibliography on randomized response: 1965-1987. *Survey Methodology*, 14, 331-346.
- POLLOCK, K.H., and BEK, Y. (1976). A comparison of three randomized response models for quantitative data. *Journal of the American Statistical Association*, 71, 884-886.
- RAO, C.R. (1952). Some theorems on minimum variance unbiased estimation. *Sankhyā (A)*, 12, 27-42.
- RAO, J.N.K. (1975). On the foundations of survey sampling. In *A Survey of Statistical Design and Linear Models*. (Ed. J.N. Srivastava). Amsterdam: North-Holland, 489-505.
- RAO, J.N.K., and BELLHOUSE, D.R. (1978). Optimal estimation of a finite population mean under generalized random permutation models. *Journal of Statistical Planning and Inference*, 2, 125-141.
- STEM, D.E., and STEINHORST, R.K. (1984). Telephone interview and mail questionnaire applications of the randomized response model. *Journal of the American Statistical Association*, 79, 555-564.
- UMESH, U.N., and PETERSON, R.A. (1991). A critical evaluation of the randomized response method. *Sociological Methods and Research*, 20, 104-138.
- WARNER, S.L. (1965). Randomized response: a survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60, 63-69.
- WARNER, S.L. (1971). The linear randomized response model. *Journal of the American Statistical Association*, 66, 884-888.
- WARNER, S.L. (1976). Optimal randomized response models. *International Statistical Review*, 44, 205-212.
- WARNER, S.L. (1986). The omitted digit randomized response model for telephone applications. *Proceedings of the Section on Survey Research Methods, American Statistical Association*.
- WOLTER, K.M. (1985). *Introduction to Variance Estimation*, New York: Springer-Verlag.

On Efficiency of Using Distinct Respondents in a Randomized Response Survey

N.S. MANGAT, R. SINGH, S. SINGH, D.R. BELLHOUSE and H.B. KASHANI¹

ABSTRACT

It is well known that the sample mean based on the distinct sample units in simple random sampling with replacement is more efficient than the sample mean based on all units selected including repetitions (Murthy 1967, pp. 65-66). Seth and Rao (1964) showed that the mean of the distinct units is less efficient than the sample mean in sampling without replacement under the same average sampling cost. Under Warner's (1965) method of randomized response we compare simple random sampling without replacement and sampling with replacement when only the distinct number of units in the sample are considered.

KEY WORDS: Simple random sampling with and without replacement; Inferences with distinct units; Warner's technique.

1. INTRODUCTION

The randomized response (RR) technique to procure trustworthy data for estimating the proportion of the population belonging to a sensitive group was first introduced by Warner (1965). Since then many developments have taken place in this area. Recently, among others, Franklin (1989), Kuk (1990), Mangat and Singh (1990, 1991), Mangat, Singh and Singh (1992) and Mangat (1994) have suggested alternative RR procedures/estimators.

In the usual simple random sampling (SRS) with replacement (WR) surveys, it is well known that the estimator of population mean based on the distinct units is always more efficient than the mean based on all selections (Murthy 1967, pp. 65-66). Also, Seth and Rao (1964) showed that, under the same average cost to sample, sampling without replacement was more efficient than with replacement sampling using the mean of the distinct sample units. This motivated the authors to investigate whether the above observation also holds in the case of Warner's pioneer RR model which is widely used in practice for selecting the respondents in the case of a survey dealing with sensitive characters. To investigate the problem we shall consider the use of four sampling strategies.

1.1 Strategy I

According to this (Warner's) procedure, each respondent included in the sample using the SRSWR method is provided with a suitable randomization device consisting of two statements of the form: (i) "I belong to sensitive group" and (ii) "I do not belong to sensitive group", represented with probabilities p and $(1 - p)$, respectively. The respondent answers "yes" or "no" according to the

randomly selected statement and to his actual status with respect to the attribute, without revealing the statement chosen. If n' persons in the sample (including repetitions) answered "yes", Warner's estimator

$$\hat{\pi} = \frac{n'/n - 1 + p}{2p - 1}, \quad p \neq .5, \quad (1)$$

is unbiased for π and its variance is given by

$$V_1(\hat{\pi}) = \frac{\pi(1 - \pi)}{n} + \frac{p(1 - p)}{n(2p - 1)^2}. \quad (2)$$

The value of p should be chosen as close to 1 or 0 as possible without threatening the degree of co-operation by respondents.

1.2 Strategy II

A sample of n respondents is drawn from a finite population of N units using SRSWR but the information from the d distinct units in the sample, $1 \leq d \leq n$, is used in the construction of the estimator. Let d' denote the respondents reporting a "yes" answer in the interview conducted with the RR device. We then consider the following estimator for π :

$$\hat{\pi}_d = \frac{d'/d - 1 + p}{2p - 1}, \quad p \neq .5. \quad (3)$$

Conditional on d distinct units, the resulting sample is a simple random sample without replacement of size d from N units. The estimator $\hat{\pi}_d$ is, therefore, unbiased for the population π .

¹ N.S. Mangat, R. Singh and S. Singh, Punjab Agricultural University, Ludhiana-141004 (India); D.R. Bellhouse, University of Western Ontario, London, Ontario, Canada, N6A 5B7; H.B. Kashani, West Oregon State College, Monmouth, OR 97361, U.S.A.

In order to study the performance of the proposed estimator $\hat{\pi}_d$, we need its variance. We give here the expression for the conditional variance $V_2(\hat{\pi}_d)$ for a given value of d . Thus

$$V_2(\hat{\pi}_d) = \frac{N-d}{N-1} \frac{\pi(1-\pi)}{d} + \frac{p(1-p)}{d(2p-1)^2}. \quad (4)$$

If E_1 and V_1 are the expectation and variance over all values of d , then we have $V_{II}(\hat{\pi}_d) = E_1 V_2(\hat{\pi}_d) + V_1 E_2(\hat{\pi}_d)$. On using (4) one gets

$$V_{II}(\hat{\pi}_d) = \left[NE_1\left(\frac{1}{d}\right) - 1 \right] \frac{\pi(1-\pi)}{N-1} + \frac{p(1-p)}{(2p-1)^2} E_1\left(\frac{1}{d}\right) \quad (5)$$

since the second term in $V_{II}(\hat{\pi}_d)$ is zero as $E_2(\hat{\pi}_d) = \pi$.

1.3 Strategy III

The sample of n respondents is selected using SRSWOR (Kim and Flueck 1978). In this case the variance of the estimator $\hat{\pi}$ in (1) can be written by replacing d in (4) by n . Thus we have

$$V_{III}(\hat{\pi}) = \frac{N-n}{N-1} \frac{\pi(1-\pi)}{n} + \frac{p(1-p)}{n(2p-1)^2}. \quad (6)$$

1.4 Strategy IV

Here the estimator is based on a WOR simple random sample of size $E(d)$. This yields the same expected cost for both in SRSWR and SRSWOR. For this scheme the estimator will be

$$\hat{\pi}_E = \frac{d'/E(d) - 1 + p}{2p-1}, \quad p \neq .5$$

with variance

$$V_{IV}(\hat{\pi}_E) = \frac{N/E(d) - 1}{N-1} \pi(1-\pi) + \frac{p(1-p)}{E(d)(2p-1)^2}. \quad (7)$$

2. EFFICIENCY COMPARISONS

It has been shown by Korwar and Serfling (1970) that, for $n \geq 3$,

$$Q - \frac{1}{720N} < E\left(\frac{1}{d}\right) \leq Q$$

where

$$Q = \frac{1}{n} + \frac{1}{2N} + \frac{n-1}{12N^2}.$$

Let us now examine the variance expression in (5). Using Q , it is easily verified that

$$\frac{NE_1(1/d) - 1}{N-1} \leq \frac{1}{n}, \quad (8)$$

in the first term on the right of (5) but that $E_1(1/d) \geq 1/n$ in the second term on the right of (5). Thus the relative efficiency of the SRSWR estimator in (1) using repeated units with respect to the SRSWR estimator in (3) using the distinct number of units will depend on the relative sizes of π and p . This is due to the fact that the repeated units can give rise to different responses because of the randomizing device and hence can provide some additional information. A sufficient condition for the inequality $V_{II}(\hat{\pi}_d) - V_I(\hat{\pi}) < 0$ to hold is obtained by using $E_1(d) = Q$. Thus we get the condition as

$$\pi(1-\pi) > \frac{n(N-1)(6N+n-1)}{N\{6Nn-12N-n(n-1)\}} \frac{p(1-p)}{(2p-1)^2}. \quad (9)$$

The above inequality is likely to hold for values of p closer to 0 or 1, the situations in which respondent jeopardy would be of concern. For example, if $N = 100$, $n = 10$ and $p = 0.9$, the inequality (9) will hold for $0.236 \leq \pi \leq 0.764$.

Similarly, Strategy II will be inferior to Strategy I if $V_{II}(\hat{\pi}_d) - V_I(\hat{\pi}) > 0$. Using $E_1(1/d) = Q - 1/720N$ this inequality reduces to

$$\pi(1-\pi) < \frac{n(N-1)\{359N+60(n-1)\}}{N\{361Nn-720N-60n(n-1)\}} \frac{p(1-p)}{(2p-1)^2}.$$

This inequality will hold for the example considered for inequality (9) whenever either $\pi \leq 0.234$ or $\pi \geq 0.764$.

On using the Cauchy-Schwarz inequality, $E(1/d) > 1/E(d)$, as in Seth and Rao (1964) we find that $V_{II}(\hat{\pi}_d) > V_{IV}(\hat{\pi}_E)$. This implies that Strategy IV is more efficient than Strategy II.

It is trivial to note that Strategy III is more efficient than Strategy I.

We know that $E(1/d) \geq 1/n$. This means $V_{II}(\hat{\pi}_d) > V_{III}(\hat{\pi})$, implying that Strategy III is more efficient than Strategy II.

Since $1/E(d) \geq 1/n$, Strategy III is more efficient than Strategy IV.

The last pair to consider consists of Strategies I and IV. Since $E(1/d) > 1/E(d)$ for $n > 1$, on using (8) we have

$$\frac{N/E(d) - 1}{N - 1} \leq \frac{1}{n}$$

implying that in (7) and (2)

$$\frac{N - E(d)}{N - 1} \frac{\pi(1 - \pi)}{E(d)} \leq \frac{\pi(1 - \pi)}{n}$$

for $n > 1$. Also $1/E(d) \geq 1/n$. This shows that the second term of (7) on the right hand side will be more than the corresponding term of (2). Thus the relative efficiencies of Strategies I and IV depend on relative values of π and p . As a numerical illustration, if $N = 100$, $n = 10$ and $p = 0.9$ then Strategy IV will be more efficient than Strategy I for $0.18 \leq \pi \leq 0.82$.

REFERENCES

- FRANKLIN, L.A. (1989). Randomized response sampling from dichotomous populations with continuous randomization. *Survey Methodology*, 15, 225-235.
- KIM, J.-I., and FLUECK, J.A. (1978). Modifications of the randomized response technique for sampling without replacement. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 346-350.
- KORWAR, R.M., and SERFLING, R.J. (1970). On averaging over distinct units in sampling with replacement. *Annals of Mathematical Statistics*, 41, 2132-2134.
- KUK, A.Y.C. (1990). Asking sensitive questions indirectly. *Biometrika*, 77, 436-438.
- MANGAT, N.S. (1994). An improved randomized response strategy. *Journal of the Royal Statistical Society, Series B*, 56, 93-95.
- MANGAT, N.S., and SINGH, R. (1990). An alternative randomized response procedure. *Biometrika*, 77, 439-442.
- MANGAT, N.S., and SINGH, R. (1991). An alternative approach to randomized response survey. *Statistica*, anno LI, 327-332.
- MANGAT, N.S., SINGH, R., and SINGH, S. (1992). An improved unrelated question randomized response strategy. *Calcutta Statistical Association Bulletin*, 42, 277-281.
- MURTHY, M.N. (1967). *Sampling Theory and Methods*, Calcutta, India: Statistical Publishing Society.
- SETH, G.R., and RAO, J.N.K. (1964). On the comparison between simple random sampling with and without replacement. *Sankhyā (A)*, 26, 85-86.
- WARNER, S.L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60, 63-69.

Cross-sectional Weighting of Longitudinal Surveys of Individuals and Households Using the Weight Share Method

PIERRE LAVALLÉE¹

ABSTRACT

Statistical agencies are conducting increasing numbers of longitudinal surveys. Although the main output of these surveys consists of longitudinal data, most of them are also expected to produce reliable cross-sectional estimates. In surveys of individuals and households, population dynamics significantly changes household composition over time. For this reason, methods of cross-sectional estimation must be adapted to the longitudinal aspect of the sample. This paper discusses in a general context the Weight Share method, of which one application is to assign a basic weight to each individual in a household. The variance estimator associated with the Weight Share method is also presented. The weighting of a longitudinal sample is then discussed when a supplementary sample is selected to improve the cross-sectional representativeness of the sample. The paper presents as an application the Survey of Labour and Income Dynamics (SLID) introduced by Statistics Canada in 1994. This longitudinal survey covers individuals' work experience, changes in income and changes in family composition.

KEY WORDS: Weight share method; Longitudinal survey; Cross-sectional estimate; Supplementary sample.

1. INTRODUCTION

Longitudinal surveys, *i.e.* surveys that follow units over time, are steadily gaining importance within statistical agencies. Statistics Canada is currently developing three major longitudinal surveys of individuals: the National Population Health Survey, the National Longitudinal Survey of Children; and the Survey of Labour and Income Dynamics (SLID).

The primary objective of these surveys is to obtain longitudinal data. One of the uses of these data is to study the changes in variables over time (*e.g.*, longitudinal data may be used to analyze the chronic aspect of poverty). A secondary objective is the production of cross-sectional estimates, in other words estimates that represent the population at a given point in time. Although these estimates are far less important than the longitudinal data, to many users they are an essential aspect of the survey. Obtaining a representative cross-sectional view of the current population constitutes a means of measuring changing situations over time. The longitudinal aspect of the survey also improves the accuracy of the measurement of change.

This paper presents an extension of the Weight Share method presented by Ernst (1989). Although the method has been developed in the context of longitudinal household surveys, it is shown that the Weight Share method can be generalized to situations where a population of interest is sampled through the use of a frame which refers to a different population, but linked somehow to the first one. In the context of longitudinal surveys, the frame can be associated to the initial population, while the population of interest can be the population a few years later. The

paper also provides a new proof of the unbiasedness of the Weight Share method together with the variance formula and variance estimator to be used with the method.

Using the Weight Share method, the question addressed in this paper is that of ensuring that the longitudinal sample can be used for cross-sectional estimation. The difficulty arises from the fact that, although the longitudinal sample remains constant, distribution of the population (individuals and households) changes over time. At the individual level, these changes are produced by such events as births and deaths, immigration and emigration, and moves within the country. Obviously, the birth or death of an individual also changes household composition; and such events as marriage, divorce, separation, departure of a child and cohabitation, are all factors that affect population distribution within the household. If we are to obtain accurate, unbiased cross-sectional estimates based on a longitudinal sample, we need an estimation method that takes these changes into account.

Our initial topic is the presentation of the Weight Share method in a general context. Secondly, we present the sample design for SLID. This is one of the major longitudinal surveys for which the production of cross-sectional estimates from a longitudinal sample is a significant problem. The survey itself is a typical longitudinal survey of individuals and households. Thirdly, we describe the use of a supplementary sample added to the initial longitudinal sample to improve the cross-sectional representativeness. Fourthly, we present the concept of basic weights, the equivalent, as it were, of sample weights. Finally, we describe the use of the Weight Share method to calculate basic weights for all individuals interviewed in SLID.

¹ Pierre Lavallée, Social Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6.

2. THE WEIGHT SHARE METHOD IN A GENERAL CONTEXT

The Weight Share method is described in Ernst (1989) in the context of longitudinal household surveys. In the same context, Kalton and Brick (1995) discuss different weighting schemes, including the Weight Share method. Various implications of using the Weight Share method for longitudinal household surveys have been described by Gailly and Lavallée (1993).

We now present this method in a general context that can be applied to several sampling situations where the population of interest needs to be sampled through the use of a frame which refers to a different population, but is linked somehow to the first one. Note that this can be viewed as a form of Network Sampling (see Thompson 1992). For example, one can imagine the need to sample young children where the only available frame is a list of names of parents. The population of interest is really the children but we need to select a sample of parents from the frame in order to obtain the sample of children. Note that the children of a particular family can be sampled through either the father or the mother. Another example is one of business surveys where an incomplete frame of establishments is available. For each selected establishment from the frame, we wish to sample the entire set of establishments belonging to the same enterprise. The missing establishments from the frame are expected to be sampled via the establishments present on the frame.

Suppose that a sample s^A of m^A units is selected from a population U^A of M^A units using some sampling design. Let π_j^A be the selection probability of unit j . We assume $\pi_j^A > 0$ for all $j \in U^A$.

Let U^B be a population of M^B units. This population is divided into N clusters where cluster i contains M_i^B units. For example, in the context of social surveys, the clusters can be households and the units can be the persons within the households. For business surveys, the clusters can be enterprises and the units can be the establishments within the enterprises. From population U^B , we are interested in estimating the total $Y = \sum_{i=1}^N \sum_{k=1}^{M_i^B} y_{ik}$ for some characteristic y .

An important constraint that is imposed in the measurement (or interviewing) process is to consider all units within the same cluster. That is, if a unit is selected in the sample, then every unit of the cluster containing the selected unit will be interviewed. This constraint is one which often arises in surveys for two reasons: cost reductions and the need for producing estimates on clusters. Referring back to the example of social surveys, there is normally a small marginal cost for interviewing all persons within the household. On the other hand, household estimates are often of interest with respect to poverty measures, for example.

We assume that there exists a *link* (or a correspondence) between each unit j of population U^A and at least one unit k of population U^B . Also, each cluster i of U^B has at least one link with a unit j of U^A . The link is identified through an indicator variable l_{jk} where $l_{jk} = 1$ if there is a link between unit $j \in U^A$ and unit $k \in U^B$ and 0 otherwise. All units of population U^A have at least one link with population U^B , i.e., $L_j^A = \sum_{k \in U^B} l_{jk} \geq 1$ for all $j \in U^A$. However, there can be zero, one or more links for a unit k of population U^B , i.e., it is possible to have $L_k^B = \sum_{j \in U^A} l_{jk} = 0$ or $L_k^B = \sum_{j \in U^A} l_{jk} > 1$ for some $k \in U^B$. This is illustrated in Figure 1.

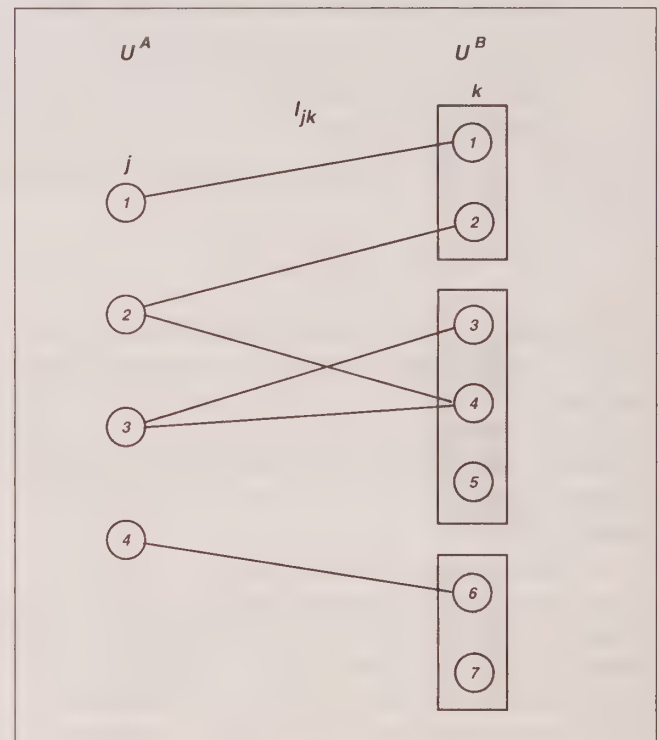


Figure 1. Links between units of populations U^A and U^B .

The estimation process presented now uses the sample s^A together with the links existing between U^A and U^B to obtain an estimation of the total Y belonging to population U^B . The links are in fact utilized as a bridge to go from population U^A to population U^B , and vice versa. Note that in practice, it might not be physically possible to directly select a sample s^B from U^B , as it has been described in the introductory examples.

To estimate the total Y , one can use the estimator

$$\hat{Y} = \sum_{i=1}^n \sum_{k=1}^{M_i^B} w_{ik} y_{ik}, \quad (1)$$

where n is the number of interviewed clusters and w_{ik} is the weight attached to unit k of cluster i . To obtain

unbiased estimates, a possible set of weights could be the inverse of the selection probabilities of the units entering into the estimator \hat{Y} . For each unit k of cluster i having a link $l_{j,ik} = 1$ with a unit j in U^A , this is possible since we have $\pi_k^B = \pi_j^A$. However, not all units of U^B necessarily have a link to U^A . Moreover, even if a link exists, it is not guaranteed that the selection probability π_j^A is known when $j \notin s^A$; the sample design used to select s^A could be, for example, a multistage sample design where the ultimate selection probability of each unit j is only known at the end of the selection process. To assign a nonzero weight w_{ik} to each unit k of cluster i entering into \hat{Y} , the Weight Share method can be used.

In general, the Weight Share method allocates to each sampled unit a basic weight established from an average of weights calculated within each cluster i entering into \hat{Y} . An *initial weight* that corresponds to the inverse of the selection probability is first obtained for unit k of cluster i of \hat{Y} having a link $l_{j,ik} = 1$ with a unit $j \in s^A$. An initial weight of zero is assigned to units not having a link. The *basic weight* is obtained by calculating the mean of the initial weights for the cluster. This weight is finally assigned to all units within the cluster. Note that the fact of allocating the same basic weight to all units has the considerable advantage of ensuring consistency of estimates for units and clusters.

Formally, each unit k of cluster i entering into \hat{Y} is assigned an initial weight w'_{ik} as follows:

$$w'_{ik} = \sum_{j=1}^{M^A} l_{j,ik} \frac{t_j}{\pi_j^A}, \quad (2)$$

where $t_j = 1$ if $j \in s^A$ and 0 otherwise. Note that a unit k having no link with any unit j of U^A automatically has an initial weight of zero.

The basic weight w_i is given by

$$w_i = \frac{\sum_{k=1}^{M_i^B} w'_{ik}}{\sum_{k=1}^{M_i^B} L_{ik}}, \quad (3)$$

where $L_{ik} = \sum_{j=1}^{M^A} l_{j,ik}$. The quantity L_{ik} represents the number of links between the units of U^A and the unit k of cluster i of population U^B . The quantity $L_i = \sum_{k=1}^{M_i^B} L_{ik}$ then corresponds to the total number of links present in cluster i .

Finally, we assign $w_{ik} = w_i$ for all $k \in i$.

2.1 Unbiasedness of the Weight Share Method

We now show that the estimator \hat{Y} with the Weight Share method is unbiased for Y . Starting with $\hat{Y} =$

$\sum_{i=1}^n w_i \sum_{k=1}^{M_i^B} y_{ik} = \sum_{i=1}^n w_i Y_i$, we replace the definition of w_i in \hat{Y} to get

$$\hat{Y} = \sum_{i=1}^n Y_i \left[\frac{\sum_{k=1}^{M_i^B} w'_{ik}}{\sum_{k=1}^{M_i^B} L_{ik}} \right] = \sum_{i=1}^n \frac{Y_i}{L_i} \sum_{k=1}^{M_i^B} w'_{ik}.$$

Letting $z_{ik} = Y_i/L_i$ for all $k \in i$, we then have

$$\hat{Y} = \sum_{i=1}^n \sum_{k=1}^{M_i^B} w'_{ik} z_{ik}. \quad (4)$$

Let a single index k be used to identify the m^B units entering into \hat{Y} ($m^B = \sum_{i=1}^n M_i^B$). By replacing w'_k by its definition (2), we obtain

$$\begin{aligned} \hat{Y} &= \sum_{k=1}^{m^B} w'_k z_k \\ &= \sum_{k=1}^{m^B} \left[\sum_{j=1}^{M^A} l_{jk} \frac{t_j}{\pi_j^A} \right] z_k. \end{aligned}$$

Now since $t_j \neq 0$ only for the units k entering into \hat{Y} , we can extend the first summation to all units k in U^B . That is,

$$\hat{Y} = \sum_{k=1}^{m^B} \left[\sum_{j=1}^{M^A} l_{jk} \frac{t_j}{\pi_j^A} \right] z_k.$$

Rearranging \hat{Y} , we finally obtain

$$\begin{aligned} \hat{Y} &= \sum_{j=1}^{M^A} \frac{t_j}{\pi_j^A} \sum_{k=1}^{m^B} l_{jk} z_k \\ &= \sum_{j=1}^{M^A} \frac{t_j}{\pi_j^A} Z_j. \end{aligned} \quad (5)$$

Now, taking the expectation gives

$$\begin{aligned} E(\hat{Y}) &= \sum_{j=1}^{M^A} \frac{E(t_j)}{\pi_j^A} Z_j \\ &= \sum_{j=1}^{M^A} Z_j = Z \end{aligned}$$

since $E(t_j) = \pi_j^A$.

It suffices now to show that $Z = Y$. First, we have

$$Z = \sum_{j=1}^{M^A} Z_j = \sum_{j=1}^{M^A} \sum_{k=1}^{M^B} l_{jk} z_k = \sum_{k=1}^{M^B} z_k \sum_{j=1}^{M^A} l_{jk}.$$

By rearranging these summations in terms of the N clusters of population U^B , we then obtain

$$\begin{aligned} Z &= \sum_{i=1}^N \sum_{k=1}^{M_i^B} z_{ik} \sum_{j=1}^{M^A} l_{j,ik} = \sum_{i=1}^N \sum_{k=1}^{M_i^B} z_{ik} L_{ik} \\ &= \sum_{i=1}^N \sum_{k=1}^{M_i^B} \frac{Y_i}{L_i} L_{ik} = \sum_{i=1}^N Y_i = Y. \end{aligned}$$

The unbiasedness of the Weight Share method can also be proved using an approach similar to the one presented by Ernst (1989).

2.2 Variance Estimation

To obtain a variance formula for \hat{Y} , we start with equation (5). Since \hat{Y} turns out to be nothing more than a Horvitz-Thompson estimator of Z (see Horvitz and Thompson 1952), the variance of \hat{Y} is then directly given by

$$\text{Var}(\hat{Y}) = \sum_{j=1}^{M^A} \sum_{j'=1}^{M^A} \frac{(\pi_{jj'}^A - \pi_j^A \pi_{j'}^A)}{\pi_j^A \pi_{j'}^A} Z_j Z_{j'}, \quad (6)$$

where $\pi_{jj'}^A$ is the joint probability of selecting units j and j' (see Särndal, Swensson and Wretman 1992 for the calculation of $\pi_{jj'}^A$ under various sampling designs).

In practice, equation (6) is simple to use. Initially, it suffices to calculate $z_k = Y_i / L_i$ for each unit $k \in i$. Then, we compute $Z_j = \sum_{k=1}^{M_j^B} l_{jk} z_k$. All that remains is to plug each Z_j into the variance equation of the Horvitz-Thompson estimator.

The variance $\text{Var}(\hat{Y})$ may be unbiasedly estimated from the following equation:

$$\widehat{\text{Var}}(\hat{Y}) = \sum_{j=1}^{m^A} \sum_{j'=1}^{m^A} \frac{(\pi_{jj'}^A - \pi_j^A \pi_{j'}^A)}{\pi_{jj'}^A \pi_j^A \pi_{j'}^A} Z_j Z_{j'}. \quad (7)$$

Another unbiased estimator of the variance $\text{Var}(\hat{Y})$ may be developed in the form of Yates and Grundy (1953). Other variance estimators are available in the literature, such as jackknife variance estimators. A jackknife variance estimator in the context of the SLID sample design is discussed in Section 3.2.3. For further details, see Wolter (1985) and Särndal, Swensson and Wretman (1992).

3. APPLICATION TO SLID

In January 1994, SLID was launched by Statistics Canada. Its aim is to observe individual activity in the labour market over time, and changes in individual income and family circumstances. To repeat, the primary aim of SLID is to provide longitudinal data. However, cross-sectional estimates will also be produced. The target population of SLID is all persons, with no distinction as to age, who live in the provinces of Canada. For operational reasons, the Territories, institutions, Indian reserves and military camps are excluded (see Lavallée 1993).

3.1 Sample Design

3.1.1 Initial Sample

The SLID longitudinal sample was drawn in January 1993 from two groups rotating out of the Canadian Labour Force Survey (LFS), making the sample a sub-sample of the LFS. The longitudinal sample for SLID is made up of close to 15,000 households. A household is defined as any person or group of persons living in a dwelling. It may consist of one person living alone, a group of people who are not related but who share the same dwelling, or it may be a family.

LFS is a continuing survey designed to produce monthly estimates of employment, self-employment and unemployment. This survey uses a stratified multi-stage design which uses an area frame in which dwellings are the final sampling units. All the individuals who are members of households that occupy the selected dwellings make up the LFS sample. In other words, LFS draws a sample of dwellings and all individuals in the households that live in the selected dwellings are interviewed. A six-group rotation plan is used to construct the sample: every month, one group that has been in the sample for six months is rotated out. Each rotation group contains approximately 10,000 households, or approximately 20,000 individuals 16 years old or more. For further details on the LFS sample plan, see Singh *et al.* (1990).

For SLID, the longitudinal sample will not be updated following its selection in January 1993. However, to give the sample some cross-sectional representativeness, *initially-absent individuals* in the population (*i.e.*, individuals who were not part of the population in the year the longitudinal sample was selected) will need to be considered in the sample in January 1994 and later. Initially-absent individuals include *newborns* (births since January 1993) and *immigrants*. Note that this addition to the sample will be cross-sectional in that only the longitudinal individuals will be permanently included in the sample.

Table 1 presents the terminology developed for SLID. After sample selection in January 93 (year 1), the population contains longitudinal individuals and initially-present individuals. In January 94 (year 2), the population contains

longitudinal individuals, initially-present individuals and initially-absent individuals. Focusing on the households containing at least one longitudinal individual (*i.e.*, *longitudinal households*), initially-present and initially-absent individuals who join these households are referred to as *cohabitants*.

Table 1
SLID Terminology

Individuals:
Longitudinal individuals: Individuals selected at year 1 in the longitudinal sample.
Initially-absent individuals: Individuals who were not part of the population in the year the longitudinal sample was selected (year 1). It includes in-migrants and newborns.
Initially-present individuals: Individuals who were part of the population of year 1 but were not selected then.
Cohabitants: Initially-absent and initially-present individuals who join a longitudinal household.
In-migrants: Individuals who, in January of year 1, were outside the ten provinces of Canada and individuals in excluded areas (the Territories, institutions, Indian reserves and military barracks).
Newborns: Births since January of year 1.
Households:
Longitudinal households: Households containing at least one longitudinal individual.

SLID will follow individual and household characteristics over time. At the time of each wave of interviews, all the members of a longitudinal household will be interviewed. The composition of the longitudinal households will change over time, as the result of a birth or the arrival of an in-migrant in the household. A part of the selection of initially-absent individuals may be based on individuals who join longitudinal households.

3.1.2 Supplementary Sample

The restriction to initially-absent individuals who join longitudinal households will unfortunately exclude households made up of initially-absent individuals only (*e.g.*, in-migrant families). To offset this shortcoming, one possibility is to draw a *Supplementary Sample*. This sample could be one of dwellings drawn directly from the ongoing LFS at each wave of interviews. Supplementary questions would then be added to the LFS questionnaire to detect households that contain *at least one in-migrant*; the households selected would then be interviewed.

Recalling that the Supplementary Sample is used for the selection of households made up solely of initially-absent individuals (*i.e.*, in-migrants and newborns), restricting this sample to in-migrants only would not cause any representativeness problem. This is because it is highly unlikely that

households containing only newborns would be found: each household normally contains at least one adult. The newborns are then already represented in the sample by the longitudinal households. Now, if the Supplementary Sample were to include newborns in addition to in-migrants, significant costs would be added to the survey. This is because the Supplementary Sample would include a complete household for each newborn selected in the Supplementary Sample, producing excessive sample growth and unnecessary costs since the newborns are already represented in the sample.

One other approach different from using the ongoing LFS could be to select the Supplementary Sample by revisiting the dwellings used for the selection of the initial sample. This method offers some practical advantages, one being the facility to go to known addresses. This approach however would bring the problem of new dwellings which were not there in January 1993. These dwellings would have a zero probability of being selected in the Supplementary Sample and a bias would therefore be introduced. This is one reason favouring the first approach, *i.e.*, detecting households that contain at least one in-migrant via the questionnaire of the ongoing LFS.

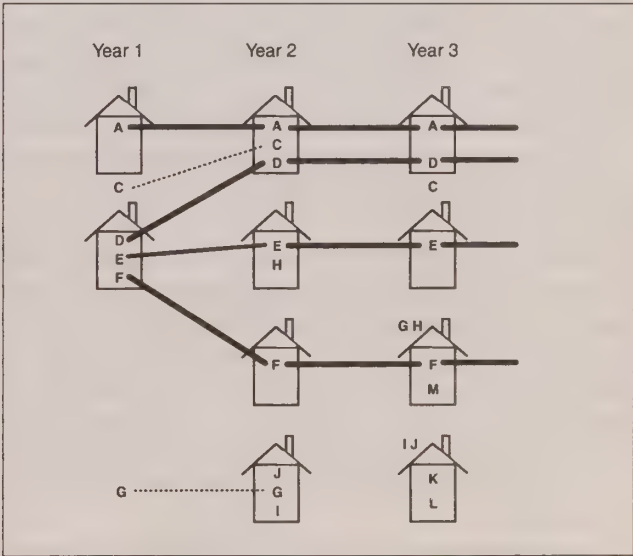


Figure 2. Selection of persons for SLID.

Figure 2 summarizes the longitudinal and cross-sectional selection of individuals. In Figure 2, the letters and houses represent individuals and households, respectively. Individuals A, D, E and F are longitudinal individuals whom we follow over time. Individual C is an initially-present individual, *i.e.*, an individual who was included in the population in year 1 but was not selected then. Initially-absent and initially-present individuals who join a longitudinal household are called cohabitants. In year 2, individual H represents an initially-absent individual who joins the sample as a

cohabitant. The fourth house in year 2 represents a household selected for the Supplementary Sample of year 2 and in which individuals I and J are initially-absent individuals (with one of the two being necessarily an in-migrant since the Supplementary Sample is restricted to them). Individual G is an initially-present individual with the same status as C. In year 3, individuals C and H have left their longitudinal households and will therefore not be interviewed. Individuals I and J who were selected in the Supplementary Sample are now replaced with the individuals of the Supplementary Sample of year 3, *i.e.*, individuals K and L. Individual M is an initially-absent individual joining a longitudinal household as a cohabitant. It may finally be noted that, for cross-sectional purposes, a selected household may contain one or more longitudinal individuals, initially-present individuals and initially-absent individuals (newborns and in-migrants).

3.2 Basic Weighting

3.2.1 General Considerations

To produce cross-sectional estimates, the longitudinal sample augmented with initially-absent individuals and initially-present individuals must be weighted. The first step is to obtain a *basic weight* for each individual in each interviewed household. The basic weight is the weight prior to adjustment or post-stratification. It is, so to speak, the equivalent of the sample weight. Note that the basic weights are useful solely for cross-sectional estimation.

The basic weights are obtained from the selection probabilities. As described above, in January 1993 (year 1), we select for SLID a sample $s^{(1)}$ of $m^{(1)}$ individuals from a population $U^{(1)}$ of $M^{(1)}$ individuals. The sample is selected through dwellings which contain households. In other words, the $m^{(1)}$ individuals are obtained by selecting $n^{(1)}$ households from $N^{(1)}$, each household I being selected with probability $\pi_I^{(1)} > 0$, $I = 1, \dots, N^{(1)}$. Let $M_I^{(1)}$ be the size of household I so that $M^{(1)} = \sum_{I=1}^{N^{(1)}} M_I^{(1)}$. Also let $\pi_j^{(1)}$ be the selection probability of individual j . This selection probability is retained throughout all waves of the survey.

For a given subsequent wave (which may be defined as year 2), the population U contains the $M^{(1)}$ individuals present at year 1, plus some $M^{(2)}$ initially-absent individuals (*i.e.*, initially absent from the population at year 1). The population of initially-absent individuals is indicated by $U^{(2)}$. Hence, the population $U = U^{(1)} \cup U^{(2)}$ contains $M = M^{(1)} + M^{(2)}$ individuals. Letting $U^{*(2)}$ be the population of $M^{*(2)}$ in-migrants of year 2, we have $U^{*(2)} \subseteq U^{(2)}$ and also $M^{*(2)} \leq M^{(2)}$. In our notation, the asterisk (*) is used to specify that the newborns have been excluded. The individuals of year 2 are contained in N households where household i is of size M_i , $i = 1, \dots, N$.

For cross-sectional representativeness, some in-migrants are selected from the Supplementary Sample. At year 2,

we then select a sample $s^{*(2)}$ of $m^{*(2)}$ individuals from the population $U^{*(2)}$ of $M^{*(2)}$ in-migrants. The Supplementary Sample is selected through households, *i.e.*, the $m^{*(2)}$ individuals are obtained by selecting $n^{*(2)}$ households. Let $\pi_j^{*(2)}$ be the selection probability of the in-migrant j . We assume $\pi_j^{*(2)} > 0$ for $j = 1, \dots, M^{*(2)}$.

One implication of selecting in-migrants through households is that other individuals (such as newborns, initially-present individuals or longitudinal individuals) can be brought in by the Supplementary Sample by living in the same household as the selected in-migrants. However, since the selection units of the Supplementary Sample are restricted to the in-migrants, these other individuals are not properly selected, say, in the Supplementary Sample, even if they will be interviewed. The selection probabilities of these individuals are in fact not well defined.

The remaining in-migrants selected for cross-sectional representativeness are those individuals who join longitudinal households, who are then considered as cohabitants. As with the newborns and initially-present individuals of the previous paragraph, the addition of cohabitants to longitudinal households brings individuals with non-well defined selection probabilities.

The individuals with non-well defined selection probabilities have entered the survey process in a “non-legitimate” way. They complicate the determination of the basic weights, as their selection probability is not well defined. In order to override this difficulty, the Weight Share method is proposed.

3.2.2 Basic Weight Calculation

The Weight Share method described in Section 2 is now applied to the SLID sample, including the Supplementary Sample. The population U^A is here represented by the union of the two distinct populations $U^{(1)}$ and $U^{*(2)}$, *i.e.*, $U^A = U^* = U^{(1)} + U^{*(2)}$. The sample s^A of $m = m^{(1)} + m^{*(2)}$ individuals corresponds to the union of the two distinct samples $s^{(1)}$ and $s^{*(2)}$. The population U^B is represented by $U = U^{(1)} + U^{(2)}$. The population $U^A = U^*$ excludes the newborns while the population $U^B = U$ includes them. The clusters of population U^B simply correspond to the N households of year 2, and hence $M_i^B = M_i$.

One possible linkage between population U^A and U^B can be established by the same individuals in populations U^A and U^B . That is, $l_{jk} = 1$ if individual j in population U^A corresponds to individual k in population U^B , and $l_{jk} = 0$ otherwise. For each individual k not being a newborn, we then have $L_{ik} = \sum_{j=1}^{M^A} l_{j,ik} = 1$. On the other hand, for each newborn k , we have $L_{ik} = \sum_{j=1}^{M^A} l_{j,ik} = 0$ since they are excluded from U^A . We now have $L_i = \sum_{k=1}^{M^B} L_{ik} = M_i^*$ where M_i^* is the size of household i excluding the newborns.

Note that this last linkage is only one among several other possibilities. One other possible linkage would be to

extend the linkage of the previous paragraph to all other persons within the household. That is, assign $l_{jk} = 1$ for all individuals k (of U^B) belonging to the same household i where individual j (of U^A) now belongs in U^B , and 0 otherwise. In other words, $l_{jk} = 1$ if individuals j and k belongs to household i . For each individual k in household i , we then have $L_{ik} = \sum_{j=1}^{M_i^A} l_{j,ik} = M_i^*$. We also get $L_i = \sum_{k=1}^{M_i^B} L_{ik} = \sum_{k=1}^{M_i^B} M_i^* = M_i^B M_i^*$. One can show that this linkage produces the same basic weighting as the one from the previous paragraph. Because the first linkage corresponds to a more natural way to link the individuals (*i.e.*, by linking only the same individuals between U^A and U^B), we will adopt the linkage proposed in the previous paragraph.

By considering the definition (2) of the initial weight w'_{ik} of individual k in household i , we obtain

$$w'_{ik} = \frac{t_{ik}^{(1)}}{\pi_{ik}^{(1)}} + \frac{t_{ik}^{*(2)}}{\pi_{ik}^{*(2)}}, \quad (8)$$

where $t_{ik}^{(1)} = 1$ if individual k is part of $s^{(1)}$ and 0 otherwise, $t_{ik}^{*(2)} = 1$ if individual k is part of $s^{*(2)}$ and 0 otherwise. This can be written more explicitly as

$$w'_{ik} = \begin{cases} 1/\pi_{ik}^{(1)} & \text{for } k \in s^{(1)} \\ 1/\pi_{ik}^{*(2)} & \text{for } k \in s^{*(2)} \\ 0 & \text{otherwise.} \end{cases} \quad (9)$$

Note that the first line of (9) corresponds to the longitudinal individuals. The second line corresponds to the in-migrants selected through the Supplementary Sample. The third line represents altogether newborns, cohabitants (if the household is a longitudinal household not part of the Supplementary Sample) and/or initially-present individuals (if the household is part of the Supplementary Sample).

The basic weight w_i of household i is obtained from

$$w_i = \frac{\sum_{k=1}^{M_i} w'_{ik}}{\sum_{k=1}^{M_i} L_{ik}} = \frac{1}{M_i^*} \sum_{k=1}^{M_i} w'_{ik}, \quad (10)$$

and finally $w_{ik} = w_i$ for $k \in i$.

Using the basic weights obtained from the Weight Share method, one can estimate the total $Y = \sum_{i=1}^N \sum_{k=1}^{M_i} y_{ik}$ of the characteristic y measured at year 2. The estimator used is the one given by equation (1). Using the definitions of the initial weights and the basic weights, \hat{Y} can be rewritten as

$$\begin{aligned} \hat{Y} &= \sum_{k=1}^{m^{(1)}} \frac{z_k^*}{\pi_k^{(1)}} + \sum_{k=1}^{m^{*(2)}} \frac{z_k^*}{\pi_k^{*(2)}} \\ &= \hat{Z}^{*(1)} + \hat{Z}^{*(2)}, \end{aligned} \quad (11)$$

where $z_k^* = \bar{Y}_i^*$ for $k \in i$ with $\bar{Y}_i^* = \sum_{k=1}^{M_i} y_{ik}/M_i^*$. Thus, estimator (11) is simply the sum of two Horvitz-Thompson estimators related to $s^{(1)}$ and $s^{*(2)}$. As shown in Section 2, this estimator is unbiased for Y .

3.2.3 Variance Estimation

The variance formula for \hat{Y} is provided by equation (6). However, assuming that the two samples $s^{(1)}$ and $s^{*(2)}$ are selected independently, we have $\text{Var}(\hat{Y}) = \text{Var}(\hat{Z}^{*(1)}) + \text{Var}(\hat{Z}^{*(2)})$, where each term has the form of equation (6). For SLID, this assumption of independance holds if the selection of the Supplementary Sample is done through LFS.

Considering $\hat{Z}^{*(1)}$, we can re-index the individuals to reflect the fact that the $m^{(1)}$ individuals were selected at year 1 through $n^{(1)}$ households. This gives

$$\begin{aligned} \hat{Z}^{*(1)} &= \sum_{k=1}^{m^{(1)}} \frac{z_k^*}{\pi_k^{(1)}} = \sum_{I=1}^{n^{(1)}} \sum_{j=1}^{M_I^{(1)}} \frac{z_{Ij}^*}{\pi_{Ij}^{(1)}} \\ &= \sum_{I=1}^{n^{(1)}} \frac{1}{\pi_I^{(1)}} \sum_{j=1}^{M_I^{(1)}} z_{Ij}^* = \sum_{I=1}^{n^{(1)}} \frac{Z_I^{*(1)}}{\pi_I^{(1)}}, \end{aligned} \quad (12)$$

since, by selecting complete households $\pi_{Ij}^{(1)} = \pi_I^{(1)}$ for $j \in I$. The variance $\text{Var}(\hat{Z}^{*(1)})$ is then directly obtained as

$$\text{Var}(\hat{Z}^{*(1)}) = \sum_{I=1}^{N^{(1)}} \sum_{I'=1}^{N^{(1)}} \frac{(\pi_{II'}^{(1)} - \pi_I^{(1)}\pi_{I'}^{(1)})}{\pi_I^{(1)}\pi_{I'}^{(1)}} Z_I^{*(1)} Z_{I'}^{*(1)}. \quad (13)$$

Considering $\hat{Z}^{*(2)}$, the individuals can also be re-indexed for consistency with $\hat{Z}^{*(1)}$, although this modification has no effect on the form of $\hat{Z}^{*(2)}$. Following the same steps used for $\text{Var}(\hat{Z}^{*(1)})$, $\text{Var}(\hat{Z}^{*(2)})$ is obtained as

$$\text{Var}(\hat{Z}^{*(2)}) = \sum_{I=1}^{N^{*(2)}} \sum_{I'=1}^{N^{*(2)}} \frac{(\pi_{II'}^{*(2)} - \pi_I^{*(2)}\pi_{I'}^{*(2)})}{\pi_I^{*(2)}\pi_{I'}^{*(2)}} Z_I^{*(2)} Z_{I'}^{*(2)}, \quad (14)$$

where $N^{*(2)}$ is the number of households of year 2 containing at least one in-migrant and $Z_I^{*(2)} = \sum_{j=1}^{M_I^{*(2)}} z_{Ij}^*$. The quantity $M_I^{*(2)}$ represents the number of in-migrants present in household I .

Finally, $\text{Var}(\hat{Y})$ is simply given by

$$\begin{aligned} \text{Var}(\hat{Y}) = & \sum_{l=1}^{N^{(1)}} \sum_{l'=1}^{N^{(1)}} \frac{(\pi_{ll'}^{(1)} - \pi_l^{(1)}\pi_{l'}^{(1)})}{\pi_l^{(1)}\pi_{l'}^{(1)}} Z_l^{*(1)} Z_{l'}^{*(1)} \\ & + \sum_{l=1}^{N^{(2)}} \sum_{l'=1}^{N^{(2)}} \frac{(\pi_{ll'}^{*(2)} - \pi_l^{*(2)}\pi_{l'}^{*(2)})}{\pi_l^{*(2)}\pi_{l'}^{*(2)}} Z_l^{*(2)} Z_{l'}^{*(2)}. \end{aligned} \quad (15)$$

The variance (15) may be unbiasedly estimated using the following equation:

$$\begin{aligned} \widehat{\text{Var}}(\hat{Y}^*) = & \sum_{l=1}^{n^{(1)}} \sum_{l'=1}^{n^{(1)}} \frac{(\pi_{ll'}^{(1)} - \pi_l^{(1)}\pi_{l'}^{(1)})}{\pi_{ll'}^{(1)}\pi_l^{(1)}\pi_{l'}^{(1)}} Z_l^{*(1)} Z_{l'}^{*(1)} \\ & + \sum_{l=1}^{n^{(2)}} \sum_{l'=1}^{n^{(2)}} \frac{(\pi_{ll'}^{*(2)} - \pi_l^{*(2)}\pi_{l'}^{*(2)})}{\pi_{ll'}^{*(2)}\pi_l^{*(2)}\pi_{l'}^{*(2)}} Z_l^{*(2)} Z_{l'}^{*(2)}. \end{aligned} \quad (16)$$

As SLID is in fact a sub-sample from LFS, the jackknife variance estimator developed for LFS (see Singh *et al.* 1990) may also be used, with minor modifications. In general, the jackknife method works as follows: the sample first is divided into random groups (or replicates, according to the LFS terminology). Then, each random group r is removed in turn from the sample and a new estimate $\hat{Y}_{(r)}$ of the total Y is computed. The different estimates $\hat{Y}_{(r)}$ are finally compared to the original estimate \hat{Y} to obtain an estimate of the variance $\text{Var}(\hat{Y})$. For further details on the jackknife method in general, see Särndal, Swensson and Wretman (1992).

Recall that LFS is based on a stratified multi-stage design which uses an area frame. Within each first-stage stratum h , the random groups (or replicates) correspond basically to the primary sampling units (PSUs). To compute the jackknife variance estimate for the estimation of the total Y , the following formula can be used:

$$\widehat{\text{Var}}_J(\hat{Y}) = \sum_h \frac{(R_h - 1)}{R_h} \sum_{r \in h} (\hat{Y}_{(hr)} - \hat{Y})^2, \quad (17)$$

where R_h is the number of replicates in stratum h and $\hat{Y}_{(hr)}$ is the estimate of Y obtained after replicate r in stratum h is removed. For LFS, both \hat{Y} and $\hat{Y}_{(hr)}$ are poststratified based on the integrated approach of Lemaître

and Dufour (1987). The plan is to use the same post-stratification approach for SLID but this discussion is out of the scope of the present paper.

ACKNOWLEDGEMENTS

The author would like to thank Carl Särndal, M.P. Singh, the Associate Editor and the referees for very useful comments which helped in improving the clarity of the paper. We would like also to gratefully acknowledge Jean-Claude Deville for suggesting the idea of an extension of the Weight Share method to a general context.

REFERENCES

- ERNST, L. (1989). Weighting issues for longitudinal household and family estimates. In *Panel Surveys*. (Eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh). New York: John Wiley and Sons, 135-159.
- GAILLY, B., and LAVALLÉE, P. (1993). Insérer des nouveaux membres dans un panel longitudinal de ménages et d'individus: simulations. CEPS/Insee, Document PSELL No. 54, Luxembourg, mai 1993.
- HORVITZ, D.G., and THOMPSON, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.
- KALTON, G., and BRICK, J.M. (1995). Weighting schemes for household panel surveys. *Survey Methodology*, 21, 33-44.
- LAVALLÉE, P. (1993). Sample representativity for the Survey of Labour and Income Dynamics. Statistics Canada, Research Paper of the Survey of Labour and Income Dynamics, Catalogue No. 93-19, December 1993.
- LEMAÎTRE, G., and DUFOUR, J. (1987). An integrated method for weighting persons and families. *Survey Methodology*, 13, 199-207.
- SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- SINGH, M.P., DREW, J.D., GAMBINO, J.G., and MAYDA, F. (1990). *Methodology of The Canadian Labour Force Survey*. Statistics Canada, Catalogue No. 71-526.
- THOMPSON, S.K. (1992). *Sampling*. New York: John Wiley and Sons.
- WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.
- YATES, F., and GRUNDY, P.M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society B*, 15, 235-261.

Weighting Schemes for Household Panel Surveys

GRAHAM KALTON and J. MICHAEL BRICK¹

ABSTRACT

Household panel surveys often start with a sample of households and then attempt to follow all the members of those households for the life of the panel. At subsequent waves data are collected for the original sample members and for all the persons who are living with the sample members at the time. It is desirable to include the data collected both for the original sample persons and for the persons living with them in making person-level cross-sectional estimates for a particular wave. Similarly, it is desirable to include data for all the households for which data are collected at a particular wave in making household-level cross-sectional estimates for that wave. This paper reviews weighting schemes that can be used for these purposes. These weighting schemes may also be used in other settings in which units have more than one way of being selected for the sample.

KEY WORDS: Cross-sectional estimates; Fair share weighting; Multiplicity weighting; Panel surveys; Weight share method.

1. INTRODUCTION

National panel surveys of household economics have been mounted in many countries in recent years. The U.S. Panel Study of Income Dynamics (PSID), conducted by the Survey Research Center of the University of Michigan, began in 1968 and has been collecting data on an annual basis since that time (Hill 1992), and the British Household Panel Survey began in 1990 (Buck *et al.* 1994). Similar household panel surveys are also in progress or are being planned in most other European countries. The U.S. Bureau of the Census started to conduct the Survey of Income and Program Participation (SIPP) in 1983 (Nelson *et al.* 1985; Kasprzyk 1988; Jabine *et al.* 1990; Citro and Kalton 1993), and Statistics Canada introduced the Survey of Labour and Income Dynamics (SLID) in 1994 (Lavallée *et al.* 1993).

A common feature to most of these household panel surveys is that they start with a national sample of households, and then follow all the members of those households for the life of the panel. Over the course of time, household compositions change in a variety of ways. Some members of original sampled households leave those households to set up on their own or to join other households, as, for example, when a daughter leaves her parental household to get married. New members may join original sampled households, as, for example, when an elderly parent moves in with the family of a child or when a child is born to a household member. In order to be able to describe the economic circumstances of sample members at different points of time, household panel surveys usually collect data

not only for the sample members but also for the individuals living with the sample members at the particular point of time. Following Lavallée (1995), these individuals are termed cohabitants in this paper. In other literature, they are often called associated persons or nonsample persons.

As the panel duration increases, the proportion of cohabitants in the sample at a wave rises. For example, in the 1984 SIPP panel, cohabitants comprise about 8.6 percent of the sample after one year and about 12.6 percent of the sample after two years (based on Table 1 in Kasprzyk and McMillen 1987). With a long-term household panel survey, the proportion of cohabitants becomes substantial after several years. The PSID, for example, defines sample members as all persons in the family units sampled in 1968 who are still alive, all the children born to these original sample members since the start of the panel, and the children of such children. In addition, the PSID collects data on the cohabitants who are living with sample members at each individual wave of data collection. Of the 20,535 individuals in interviewed family units in 1992, 41.2 percent were original sample members, 34.6 percent were the children of original sample members born since the start of the panel and children of such children, and 24.2 percent were cohabitants (excluding the Latino sample that was added in 1990) (Hill 1995).

This paper reviews methods of weighting the data collected from both sample persons and cohabitants in order to produce unbiased (or approximately unbiased) estimates of population parameters. In considering the analysis of a household panel survey, three different types of analysis may usefully be distinguished:

¹ Graham Kalton and J. Michael Brick, Westat Inc., 1650 Research Blvd., Rockville, MD 20850, U.S.A.

- Cross-sectional analyses of households at a particular point in time;
- Cross-sectional analyses of individuals at a particular point in time;
- Longitudinal analyses of individuals over a period of time.

Weighting schemes for these three types of analysis are discussed in later sections. Longitudinal analyses of households over a period of time are not treated here because of the problematic nature of this type of analysis caused by changes in household composition (see, for example, Duncan and Hill 1985).

The weighting schemes used in household panel surveys need to account for the fact that households and individuals included in the survey at a particular wave may have more than one route by which they can be selected. At a given wave a household and its members are included in the sample if any of the original households (*i.e.*, households existing at the time of the initial selection) from which the current household has drawn members was selected. With the usual weighting approach, households are assigned weights inversely proportional to their joint selection probabilities, taking account of the different ways they can be selected. However, this approach cannot be applied with most household panel surveys because these joint selection probabilities cannot be determined. The alternative weighting approach reviewed here, termed by Lavallée (1995) the *weight share method*, avoids the need to know the joint selection probabilities of sample elements, but it introduces a random variation into the weights. Since this random variation results in a loss in precision of the survey estimates as compared with the inverse selection probability weighting scheme, this alternative approach should be considered only for situations where the joint selection probabilities cannot be ascertained. This situation often applies in household panel surveys and also in a number of other sample designs where elements can be selected by different routes.

In order to prepare for the discussion of weighting schemes for household panel surveys, the next section elaborates on the household changes that can occur over time, and the types of individuals involved. Sections 3, 4 and 5 then discuss weighting schemes that may be used for the three different forms of analysis described above. These sections deal with weighting schemes for unequal selection probabilities, without the complications of adjustments for nonresponse and noncoverage. The discussion relies heavily on previous work by Ernst (1989), Gailly and Lavallée (1993), Huang (1984), Judkins *et al.* (1984), Lavallée and Hunter (1992), and Little (1989). Section 6 then briefly reviews the issues involved in making adjustments to the weights to compensate for missing data arising from nonresponse and noncoverage. Section 7 presents some concluding remarks, and provides an illustration of another application of the weight share method.

2. CHANGES IN POPULATION AND HOUSEHOLD COMPOSITION OVER TIME

In analyzing a panel survey, it needs to be recognized that survey populations change over time. With household panel surveys it is important to distinguish between changes in population composition and changes in household composition.

The composition of a survey population changes over time because some individuals leave the population, some enter the population, and some may leave and join the population more than once. Individuals leave the population through death, emigration, or entering an institution (for surveys of the noninstitutional population). They enter the population through birth (or reaching the specified minimum age), immigration, and leaving an institution.

Households change composition over time for many different reasons, including deaths, births, marriages and divorces. For example, a household at time 1 may contain several individuals who end up in a number of different households at time 2. These individuals may set up new households on their own, they may join individuals who were in one or more households at time 1, or they may join individuals who were not in the population at time 1. One or more of the individuals may leave the population during the intervening period.

Consider a simple sample design in which households are selected independently at time 1 with equal probability. At time 2, the sample of households comprises all the households that contain one or more individuals from the households sampled at time 1, and the sample of individuals at time 2 comprises all the members of the sampled households at time 2. The samples of households and individuals at time 2 are selected with unequal probabilities. For instance, the selection probability of a household at time 2 that contains individuals from three households at time 1 is three times greater than that of a household at time 2 that contains individuals from only one household at time 1. Similarly, the individuals in that household have three times the probability of selection. Thus weighting schemes that compensate for these unequal selection probabilities are needed for the analysis of the resultant data.

Changes in population composition occur when individuals leave or enter the population. An individual sampled at time 1 who leaves the population before time 2 reduces the sample size for time 2 but does not otherwise affect cross-sectional estimates at time 2. In essence, the sampling frame for the time 2 population is the time 1 population, with the leavers in the intervening period being treated as blanks on the frame. Simply omitting the selected blanks from the time 2 sample causes no bias in the survey estimates (see, for example, Kish 1965). The situation with regard to entrants is, however, less straightforward. The household panel survey enumeration rule described above

incorporates new entrants who join households that contain individuals who were eligible for the initial sample into the population for cross-sectional estimates for later time points. However, new entrants who set up their own households are not represented in person-level analyses at later waves of the panel. Equally, households composed of only new entrants are not represented in household-level analyses at later waves.

The failure of household panel surveys to cover households composed of only new entrants presents a problem for cross-sectional analyses of later waves of the panel. If these households and their members constitute a negligible proportion of the population, the solution may be to simply ignore the problem. However, if the proportion is appreciable, as can occur in later waves of a long-term panel, alternative solutions may be called for. One possibility is to add a supplementary sample of new entrants (e.g., immigrants) to the panel, as discussed by Lavallée (1995) for the SLID. This solution is, however, often impracticable. Another solution is to limit the population of inference to persons who were members of the population at the start of the panel. New entrants found living with sample members are then excluded from the sample. This solution provides a clearcut definition of the population of inference. Whether the solution is appropriate depends on whether that definition can adequately satisfy the survey objectives.

Changes in population composition pose problems for longitudinal analyses of individuals. For many purposes, the population of inference is restricted to those who were present in the population throughout the time period of observation specified for the analysis. The inclusion of cohabitants in longitudinal analysis also creates problems. If the time period for the longitudinal analysis starts at the beginning of the panel, the analysis can be restricted straightforwardly to original sample members. If the time period starts later, it is tempting to include both original sample members and cohabitants joining the panel before the start of the analytic time period. However, the usual enumeration rules for household panel surveys specify that data are collected for cohabitants only while they continue to live with original sample members, that is, they are not followed if they cease to live with such persons. Unless the time period is short enough that the number of cohabitants who cease to live with sample persons in that period is negligible, this enumeration rule makes it problematic to include cohabitants in longitudinal analyses. This problem is discussed further in Section 5.

3. CROSS-SECTIONAL ESTIMATES FOR HOUSEHOLDS

This section considers weighting schemes that may be used to produce cross-sectional estimates for households for any wave of a household panel survey after the first.

At the first wave a sample of households is selected and all the individuals in the sampled households become panel members to be followed throughout the life of the panel or until they leave the survey population. At a subsequent wave, wave t , the household sample comprises all the households in which panel members reside. Households that consist of new entrants only are not represented in the sample at later waves. Such households are ignored here. Complications of nonresponse are deferred until Section 6.

Consider the estimation of the total Y for all H households in the population at time t :

$$Y = \sum_{i=1}^H Y_i. \quad (3.1)$$

A general estimator for this total can be expressed as

$$\hat{Y} = \sum_{i=1}^H w_i Y_i,$$

where w_i is a random variable that takes the value $w_i = 0$ if household i is not in the sample. The expectation of \hat{Y} is

$$E(\hat{Y}) = \sum_{i=1}^H E(w_i) Y_i. \quad (3.2)$$

By comparing equations (3.1) and (3.2), it can be seen that \hat{Y} is unbiased for Y for any weighting scheme for which $E(w_i) = 1$ for all i .

There are many ways to satisfy the condition $E(w_i) = 1$. Three will be treated here. First, consider a standard *inverse selection probability weighting scheme*. The probability of a household being in the sample at time t is the probability of one or more of the households at time 1 from which it has drawn members being selected for the original sample. The probability of household H_i being in the sample at time t is then

$$\begin{aligned} P(H_i) &= P(h_j \cup h_k \cup h_t \cup \dots) \\ &= \sum p_j - \sum \sum p_{jk} + \sum \sum \sum p_{jkl} - \dots, \end{aligned} \quad (3.3)$$

where $P(h_j \cup h_k \cup h_t \cup \dots)$ is the selection probability of the union of original households h_j , h_k , h_t , etc. for the original sample, p_j is the selection probability of original household h_j for the original sample, p_{jk} is the joint selection probability of original households h_j and h_k for the original sample, etc. and where households h_j , h_k , h_t , etc. each contain at least one member who is currently in household H_i . The weight for each sampled household is then $w_i = 1/P(H_i)$. With this weighting scheme,

$$E(w_i) = P(H_i) [1/P(H_i)] + [1 - P(H_i)] 0 = 1,$$

satisfying the condition for an unbiased estimator of a population total.

In practice, the computation $P(H_i)$ will generally not be as complex as equation (3.3) might suggest because the number of original households represented in household H_i is usually small. With, say, two original households involved, $P(H_i)$ reduces to

$$P(H_i) = P(h_1 \cup h_2) = p_1 + p_2 - p_{12}. \quad (3.4)$$

A problem with the application of the inverse selection probability approach is that p_j may be known only for households selected for the original sample, and not for other households. Also the joint probability may not be known. Even when the original sample was selected with equal probabilities, so that all the p_j are the same, the joint probability may depend on the sample design (for instance, whether the two households were in the same segment or not). The difficulty of obtaining $P(H_i)$ is a major drawback with the inverse selection probability approach.

An alternative strategy for developing the weights for time t is to base them only on the selection probabilities of households selected for the original sample, thus avoiding the difficulty in obtaining $P(H_i)$ noted above. One approach is to identify the set of households h_j at time 1 that would result in household H_i being in the sample at time t , and compute the weight for household H_i as

$$w_i = \sum_j \alpha_{ij} w'_{ij}, \quad (3.5)$$

where $w'_{ij} = 1/p_j$ if household h_j , which has at least one member in household H_i , was selected for the original sample and $w'_{ij} = 0$ if not, and where α_{ij} are any set of constants satisfying $\sum_j \alpha_{ij} = 1$.

With this approach,

$$E(w'_{ij}) = p_j(1/p_j) + (1 - p_j)0 = 1,$$

and hence

$$E(w_i) = \sum_j \alpha_{ij} = 1.$$

Thus, the use of weights w_i will yield unbiased estimators of totals for the household population for any choice of constants α_{ij} , provided that $\sum_j \alpha_{ij} = 1$. As indicated above, the principal advantage of this type of scheme is that it requires information only on the initial selection probabilities of the original households that were sampled at time 1, which are known. It does not require information on the initial selection probabilities of the other original households that have members in the current household, which are often not known.

A natural choice of α_{ij} is to make them equal for all the original households that lead to the selection of household H_i at time t . Huang (1984) terms this scheme a

multiplicity approach. Here the scheme will be called an *equal household weighting scheme*. With this scheme

$$w_i = \sum_j w'_{ij}/C_i, \quad (3.6)$$

where C_i is the number of original households represented in household H_i at time t .

An alternative version of the above approach is one based on original sample persons rather than households. In this case, let I_{ijk} denote individual k from original household j in household i . Then

$$w_i = \sum_j \sum_k \alpha_{ijk} w'_{ijk},$$

where $w'_{ijk} = 1/p_j$ if individual k in household h_j was in the original sample and $w'_{ijk} = 0$ if not, and where the α_{ijk} are any set of constants satisfying $\sum_j \sum_k \alpha_{ijk} = 1$. Since the probability of an individual being selected for the original sample is the same as that of that individual's household,

$$E(w'_{ijk}) = p_j(1/p_j) + (1 - p_j)0 = 1.$$

In this case, the natural choice of the constants α_{ijk} is to make them equal for all members of the current household who were eligible for selection for the original sample. This produces what has been termed the fair share weighting scheme (Huang 1984; Ernst 1989). This scheme is termed here an *equal person weighting scheme*. With this scheme

$$w_i = \frac{1}{M_i} \sum_j M_{ij} w'_{ij},$$

where $w'_{ij} = w'_{ijk}$ is constant for all individuals in household H_i emanating from the same sampled household at time 1, M_{ij} is the number of individuals in household H_i coming from household h_j , and $M_i = \sum_j M_{ij}$ is the number of individuals in household H_i who were eligible for the sample at time 1. The equal person weighting scheme is applied in the SIPP and is proposed for use in the SLID.

Although developed here in terms of persons rather than households, it is readily apparent that the equal person weighting scheme could equally have been generated in terms of households. As shown above, the household weight $w_i = \sum_j \alpha_{ij} w'_{ij}$ satisfies the condition $E(w_i) = 1$ for any set of constants α_{ij} such that $\sum_j \alpha_{ij} = 1$. The equal household weighting scheme chooses $\alpha_{ij} = 1/C_i$, with $\sum_j \alpha_{ij} = 1$. The choice $\alpha_{ij} = M_{ij}/M_i$, with $\sum_j \alpha_{ij} = 1$, leads to the equal person weighting scheme.

It is instructive to compare the inverse selection probability weighting scheme with the equal household and equal person weighting schemes in a simple case. Following Little (1989), consider household H_i selected at time t

with household members coming from two original households. Let p_1 and p_2 denote the selection probabilities for the original households, and let p_{12} denote their joint selection probability. Under the inverse selection probability approach, the household weight is

$$w_i^* = \frac{1}{p_1 + p_2 - p_{12}},$$

as indicated above.

Under the equal person weighting scheme the weight for household H_i depends on which household or households were selected for the original sample:

$w_i = P_1/p_1$ if only household h_1 was selected;
 $w_i = P_2/p_2$ if only household h_2 was selected;
 $w_i = (P_1/p_1) + (P_2/p_2)$ if both h_1 and h_2 were selected;

where P_1 and P_2 are the proportions of members of household H_i who came from households h_1 and h_2 , respectively (excluding any new entrants to the population). The probability of only household h_1 being selected is $(p_1 - p_{12})$, of only household h_2 being selected is $(p_2 - p_{12})$, and of both households being selected is p_{12} . The expected value of the weight conditional on household H_i being in the sample is thus

$$E(w_i | H_i \text{ in sample}) = \frac{(p_1 - p_{12})(P_1/p_1) + (p_2 - p_{12})(P_2/p_2) + p_{12}[(P_1/p_1) + (P_2/p_2)]}{p_1 + p_2 - p_{12}},$$

i.e.,

$$E(w_i | H_i \text{ in sample}) = \frac{1}{p_1 + p_2 - p_{12}} = w_i^*.$$

As this result demonstrates, the weight for household H_i varies depending on which original households were selected, but in expectation the weight is the same as that obtained from the inverse selection probability approach.

Results for the expectation of the weight of household H_i under the equal household weighting scheme can be readily obtained as a special case of the above derivation in which $P_1 = P_2 = 1/2$. In expectation, the weight is the same as that for the inverse selection probability approach.

Given that the weight $w_i = \sum_j \alpha_{ij} w'_{ij}$ satisfies the condition $E(w_i) = 1$ for any set of α_{ij} such that $\sum_j \alpha_{ij} = 1$, the question arises as to the optimal choice of the α_{ij} . One approach is to choose the α_{ij} to minimize the variance of the estimated total \hat{Y} .

The variance of \hat{Y} may be expressed as

$$V(\hat{Y}) = VE(\hat{Y} | s) + EV(\hat{Y} | s), \quad (3.7)$$

where s denotes the set of households in the sample at time t . Now

$$\begin{aligned} E(\hat{Y} | s) &= E\left(\sum_{i=1}^H w_i Y_i | s\right) \\ &= \sum_{i=1}^s E(w_i | H_i) Y_i = \sum_{i=1}^s w_i^* Y_i = \hat{Y}^*, \end{aligned}$$

where \hat{Y}^* is the standard inverse selection probability estimator. Thus

$$VE(\hat{Y} | s) = V(\hat{Y}^*).$$

The first term in equation (3.7) is thus the variance of the standard inverse selection probability estimator, and the second term is the additional variance resulting from the use of weighting schemes from the class (3.5), $w_i = \sum_j \alpha_{ij} w'_{ij}$. The α_{ij} may then be chosen to minimize $EV(\hat{Y} | s)$.

Consider

$$\begin{aligned} V(\hat{Y} | s) &= V\left(\sum_{i=1}^H w_i Y_i | s\right) \\ &= \sum_{i=1}^s Y_i^2 V(w_i | H_i) + \\ &\quad \sum_{i \neq i'} \sum Y_i Y_{i'} \text{Cov}(w_i, w_{i'} | H_i, H_{i'}). \end{aligned}$$

Assuming $\text{Cov}(w_i, w_{i'} | H_i, H_{i'}) = 0$,

$$\begin{aligned} V(\hat{Y} | s) &= \sum Y_i^2 V(w_i | H_i) \\ &= \sum Y_i^2 [E(w_i^2 | H_i) - w_i^{*2}], \end{aligned}$$

since, as noted above, $E(w_i | H_i) = w_i^*$. Thus, assuming $\text{Cov}(w_i, w_{i'} | H_i, H_{i'}) = 0$, $V(\hat{Y} | s)$ is minimized when $E(w_i^2 | H_i)$ is minimized.

Consider the application of this approach to the simple case discussed above in which H_i is composed of members from two original households and let $w_i = \alpha_i w'_{i1} + (1 - \alpha_i) w'_{i2}$. Then

$$\begin{aligned} E(w_i^2 | H_i) &= \\ \frac{(p_1 - p_{12}) \frac{\alpha_i^2}{p_1^2} + (p_2 - p_{12}) \frac{(1 - \alpha_i)^2}{p_2^2} + p_{12} \left(\frac{\alpha_i}{p_1} + \frac{1 - \alpha_i}{p_2} \right)^2}{p_1 + p_2 - p_{12}}. \end{aligned}$$

Minimizing $E(w_i^2 | H_i)$ is equivalent to minimizing

$$\begin{aligned} \Delta &= (p_1 - p_{12}) p_2^2 \alpha_i^2 + (p_2 - p_{12}) p_1^2 (1 - \alpha_i)^2 \\ &\quad + p_{12} [(p_2 - p_1) \alpha_i + p_1]^2. \end{aligned}$$

Then

$$\frac{\partial \Delta}{\partial \alpha_i} = 2(p_1 - p_{12})p_2^2\alpha_i - 2(p_2 - p_{12})p_1^2(1 - \alpha_i) + 2p_{12}(p_2 - p_1)[(p_2 - p_1)\alpha_i + p_1].$$

Solving $\partial \Delta / \partial \alpha_i = 0$ for α_i gives the optimum α_i as

$$\alpha_{oi} = \left(1 + \frac{p_2 - p_{12}}{p_1 - p_{12}}\right)^{-1}. \quad (3.8)$$

If the original households are selected independently, *i.e.*, $p_{12} = p_1 p_2$,

$$\alpha_{oi} = \left[1 + \frac{p_2(1 - p_1)}{p_1(1 - p_2)}\right]^{-1} = \left[1 + \frac{\psi_2}{\psi_1}\right]^{-1}, \quad (3.9)$$

where $\psi_j = p_j / (1 - p_j)$ is the odds of original household h_j being selected.

Irrespective of whether the households are sampled independently, in the special case of an equal probability (epsem) sample of households initially, with $p_1 = p_2$,

$$\alpha_{oi} = \frac{1}{2}.$$

Thus, in the two-household case, the equal household weighting scheme minimizes the variance of the household weights around the inverse selection probability weight when the initial sample is an epsem one.

The optimal choice of α_{oi} given by (3.8) requires knowledge of p_1 , p_2 and p_{12} , and that given by (3.9) requires independence and knowledge of p_1 and p_2 . If these probabilities were known, then the standard inverse selection probability weight could be employed and would be preferable. In the case of an approximately epsem sample, the equal household weighting scheme should be close to the optimal, at least for the case where the members of the household at time t come from one or two households at the initial wave. This would apply, for instance, in the case of an epsem initial sample, with perhaps a few departures from epsem. With the equal household weighting scheme, when only one of the C_i original households, h_j , represented in H_i was selected for the original sample (as will generally be the case), then the weight for H_i is simply $1/C_i p_j$.

In the case of a non-epsem initial sample, the choice of the α_{ij} would ideally depend on the original household selection probabilities. However, since these probabilities are unknown, that approach cannot be applied. By default, the equal household or equal person weighting schemes may therefore be employed in this case. The use of these schemes (or any scheme with constant α_{ij} 's satisfying $\sum_j \alpha_{ij} = 1$) with a non-epsem initial sample still results in

an unbiased estimate \hat{Y} . The drawback to these schemes in such a case is only that the α_{ij} are suboptimal in terms of minimizing the variance of \hat{Y} .

It should be noted that the equal household weighting scheme requires information on the number of original households h_j contributing members to household H_i at time t . That number may be difficult to determine in some cases. Consider, for example, a household at time t that contains two cohabitants. It may sometimes be difficult to determine whether these two persons were in a single household or in two separate households at the time of the initial sample selection. The equal person weighting scheme has the attractive feature of avoiding the need for Wave 1 household information, except for persons in sampled households at Wave 1. This feature provides an important reason for preferring the equal person to the equal household weighting scheme.

4. CROSS-SECTIONAL ESTIMATES FOR INDIVIDUALS

In producing cross-sectional estimates for individuals for any wave of a household panel survey after the first, it needs to be recognized that some new entrants will have joined the survey population since the start of the panel. New entrants who join households that contain one or more members of the original population can be represented in cross-sectional estimates for later waves, but new entrants living in households that do not contain any members of the original population are not covered (unless a special sample of them can be taken). The former type of new entrants is included in the weighting procedure described below, but the latter type is not.

Let there be N individuals in the population at time t , with N_i individuals in household H_i ($i = 1, 2, \dots, H$) and $\sum N_i = N$. The members of household H_i come from households h_j, h_k, h_l , etc., at time 1. Let M_{ij} denote the number of members of household H_i at time t who were in household h_j at the start of the panel. The sum $M = \sum \sum M_{ij}$ is less than the population size at time 1 because of leavers from the population in the period from time 1 to time t , and $M < N$ because of new entrants to the population who are in households containing members from the original population.

Consider now the estimation of a total for the population of individuals at time t :

$$Y = \sum_{i=1}^H \sum_{k=1}^{N_i} Y_{ik}. \quad (4.1)$$

where Y_{ik} is the value for individual k in household H_i . As in the household case discussed in the previous section, a general estimator for this total can be expressed as

$$\hat{Y} = \sum_{i=1}^H \sum_{k=1}^{N_i} w_{ik} Y_{ik}, \quad (4.2)$$

where w_{ik} is a random variable that takes the value $w_{ik} = 0$ if individual k in household H_i is not in the sample. The estimator \hat{Y} is unbiased for Y provided that $E(w_{ik}) = 1$ for all i and k .

As noted earlier, there are many ways to satisfy the condition $E(w_{ik}) = 1$. It is instructive to consider three of them. First, let $w_{ik} = 0$ for all individuals not in the original sample. In this case, the estimator \hat{Y} discards cohabitants. Let p_{ik} denote the probability of a member of the original population, individual k residing in household H_i at time t , being selected for the initial sample, and let $w_{ik} = 1/p_{ik}$. Then, for such an individual

$$E(w_{ik}) = p_{ik}(1/p_{ik}) + (1 - p_{ik})0 = 1.$$

With this scheme, all new entrants to the population have $w_{ik} = 0$ with certainty. Thus \hat{Y} in (4.2) provides an unbiased estimator of the total for the original population that is still present at time t , but does not include a component for the new entrants.

Modifications to the above procedure can be made to cover certain types of new entrants. For instance, births to sampled mothers can be included by assigning them the weights of their mothers, or if, as in the SIPP, the survey population is taken to be adults aged 16 and over, those under 16 at the start of the panel can be treated as sampled persons with assigned probabilities, and they can be included in the analyses of later waves after they have attained the age of 16. Such modifications do not, however, handle all types of new entrants. Provided that the proportion of other types of new entrants is small, this deficiency may not be a serious concern.

The weighting scheme that restricts the analysis to original sample persons, plus certain specified new entrants, is employed with the PSID. Its limitation is that it fails to make direct use of data collected for cohabitants. Such data may be used to provide information on the situation of sample persons, but the cohabitants are excluded from the sample for the analysis.

In order to include cohabitants in cross-sectional analyses for time t they need to be assigned positive weights. Noting that the probability of an individual being selected for the sample is the same as that of his or her household, weighting schemes for cross-sectional analyses of individuals at wave t can be obtained directly from those for households given in Section 3. Here we will develop the general strategy of producing weights for cross-sectional analysis at time t based only on the selection probabilities of members of the original sample, thus avoiding the problems with the inverse selection probability approach noted in Section 3.

Let I_{ijk} denote individual k from original household h_j who is now in household H_i . Let w_i denote the weight for every member of household H_i for cross-sectional analyses at time t , and let

$$w_i = \sum_j \sum_k \alpha_{ijk} w'_{ijk}$$

where $w'_{ijk} = 1/p_j$ if household h_j was in the original sample and $w'_{ijk} = 0$ if not. Then, as before, $E(w'_{ijk}) = 1$ for members of the original population. New entrants, for whom $p_j = 0$, may be handled by setting $\alpha_{ijk} = 0$. Then

$$E(w_i) = \sum_j \sum_k \alpha_{ijk} E(w'_{ijk}) = \sum_j \sum_k \alpha_{ijk} = 1$$

provided that $\sum_j \sum_k \alpha_{ijk} = 1$. Under this condition \hat{Y} is unbiased for Y .

A natural choice of α_{ijk} is to set $\alpha_{ijk} = 1/M_i$ for all members of the original population. This is the equal person weighting scheme in which every member of household H_i at time t (including new entrants) receives the weight

$$w_i = \sum_j \sum_k w'_{ijk} / M_i.$$

Another choice of the α_{ijk} is that used for the equal household weighting scheme. Let C_i denote the number of original households that have members in household H_i at time t . Then $\sum_j \sum_k \alpha_{ijk} = 1$ can be divided equally between households, with each member of original household h_j being assigned a value of $\alpha_{ijk} = 1/C_i M_{ij}$. Then for original household h_j

$$\sum_k \alpha_{ijk} = 1/C_i.$$

The derivation of the α_{ijk} to minimize the variance of the estimated total \hat{Y} for the population of individuals follows directly from the corresponding derivation for the population of households given in Section 3. The estimated total for the population of individuals is

$$\hat{Y} = \sum_i^s \sum_k^{N_i} w_{ik} Y_{ik} = \sum_i^s \sum_k^{N_i} w_i Y_{ik},$$

since the weights for every individual in sampled household H_i are the same. This estimated total can be expressed as

$$\hat{Y} = \sum_i^s w_i Y_i,$$

where $Y_i = \sum_k Y_{ik}$ is the household total for H_i . Thus \hat{Y} can be expressed as a household total, and the results of Section 3 can be applied directly.

Consider the example from Section 3 in which H_i is composed of members from only two original households, perhaps together with one or more new entrants. In this case the person-level weight $w_i = \sum_j \sum_k \alpha_{ijk} w'_{ijk}$ reduces to

$$\begin{aligned} w_i &= \left(\sum_k \alpha_{i1k} \right) w'_{i1} + \left(\sum_k \alpha_{i2k} \right) w'_{i2} \\ &= \alpha_i w'_{i1} + (1 - \alpha_i) w'_{i2}, \end{aligned}$$

where $\alpha_i = \sum_k \alpha_{i1k}$. As shown in equation (3.8), the optimum value of α_i is

$$\alpha_{oi} = \left(1 + \frac{p_2 - p_{12}}{p_1 - p_{12}} \right)^{-1}.$$

The individual values α_{ijk} are not needed for computing the w_i ; only the original household totals $\sum_k \alpha_{ijk}$ are required. If individual values are needed for the α_{ijk} , they may be simply assigned as $\sum_k \alpha_{ijk} / M_{ij}$.

As in the household case, the optimum weighting α_{oi} requires knowledge of p_1 , p_2 and p_{12} . If these probabilities are known, the standard inverse selection probability weight w_i^* can be computed, and would be preferred. In the case of an approximately epsem sample, the equal household weighting scheme should fare well. However, the equal household weighting scheme requires information on the number of original households contributing members to current household H_i , and this information may not always be available. As discussed in Section 3, for this reason the equal person weighting scheme may be preferred.

5. LONGITUDINAL ANALYSES OF INDIVIDUALS

A key analytic advantage of a panel survey is the ability to conduct longitudinal analyses relating variables for the same sampled units measured at different time points. Since all persons in original sampled households are followed throughout the life of the panel or until they leave the survey population, the data they provide may be readily analyzed longitudinally for any time period within the panel's time span (although nonresponse adjustments may be needed for panel attrition). Thus, for example, in a ten-year panel, data for original sampled persons may be analyzed from year 1 to year 10, from year 5 to year 9, or for any other period. New entrants (e.g., births) may be included in the analysis for periods beginning after the start of the panel provided that they are treated as panel members who are followed throughout the panel even when they leave the households of original sampled persons.

Given the weighting schemes described in the previous section, cohabitants can be included in cross-sectional analyses of later waves. These weighting schemes provide a cross-sectional representation of the population at any wave of the panel (apart from new entrants not living with original population members). It is then possible to consider all the sample of original sample members and cohabitants at time t as the initial sample of a new panel that may be used for longitudinal analyses from time t to $(t + k)$. This procedure is, for instance, used in the SIPP, where all original sample members and cohabitants present at the start of the second year of the panel are included in analyses relating to that year.

The limitation to the inclusion of cohabitants in longitudinal analysis is that the following rules used in most household panel surveys specify that cohabitants are dropped from the panel if they cease living with original sample persons. Thus, cohabitants who live with original sample members at the start of the analysis period but who cease to live with them before the end of that period effectively become nonrespondents. If the analysis period is relatively short, the number of such nonrespondents may be small and the risk of serious nonresponse bias may be negligible. If the analysis period is a long one, however, the number of not-followed cohabitants may be appreciable, causing concerns about potential bias. The issue here is one of a trade-off between the reduced variance due to the increase in sample size from including cohabitants in the analysis versus the increased bias resulting from the additional nonresponse caused by failing to follow cohabitants leaving the households of original sample persons.

The additional nonresponse bias can be avoided by changing the following rules to specify that cohabitants are to be followed from the time they join the panel for the rest of the life of the panel, or until they leave the survey population, irrespective of whether they continue to live with original sample members. This change, however, leads to an expanding panel and the need for additional resources. Not only do data need to be collected for cohabitants at waves after they cease to live with sample persons, but data also need to be collected for any persons with whom the cohabitants live at later waves.

6. ADJUSTMENTS TO COMPENSATE FOR NONRESPONSE AND NONCOVERAGE

The discussion thus far has assumed that data are collected for all sampled persons and their cohabitants and that all the target population is covered by the sampling procedures. In practice both these assumptions are violated. Nonresponse is present in nearly all surveys and is of particular concern in household panel surveys, where some

sampled households fail to respond at the initial wave and others fail to respond at some of the subsequent waves. The sampling frames used in most surveys are subject to some degree of noncoverage, and in later waves of household panel surveys there is an additional source of noncoverage associated with new entrants to the population who are not living with members of the original population.

In a simple cross-sectional survey, missing data can be classified into item nonresponse, total nonresponse and noncoverage. Imputation procedures can then be used to assign values for item nonresponses, weighting adjustments can be applied to compensate for total nonresponse, and poststratification adjustments can be applied to compensate for nonresponse and noncoverage. The situation is made far more complex in panel surveys by the occurrence of wave nonresponse, which arises when a sampled element responds for some but not all of the waves for which it was eligible. Not only do methods need to be devised to compensate for wave nonresponse, but also the preferred methods of compensation may depend on the type of analysis to be performed, in particular whether cross-sectional or longitudinal analyses are to be conducted.

From one perspective wave nonresponse can be viewed as a set of item nonresponses in the element's longitudinal record, suggesting that imputation may be used to fill in the missing values. Alternatively, it can be treated as total nonresponse, handled by weighting adjustments. The imputation approach is more natural for the creation of a panel file for longitudinal analysis, whereas the weighting approach is more natural for the creation of a cross-sectional file for the analysis of the data collected at a single wave.

The attraction of the imputation approach with a longitudinal file is that it retains all the reported data, whereas the weighting approach discards the reported data for all the elements that fail to provide data for one or more waves for which they were eligible. However, the imputation approach may involve the fabrication of a large amount of data, especially when an element fails to respond at several waves. Thus, for panel files, a compromise solution may be preferred, imputing responses for elements with few missing waves and using weighting adjustments to compensate for those with several missing waves (including total nonrespondents). In the SIPP, for example, imputation is used to assign responses for sample persons with a single missing wave that is bounded on both sides by responding waves, and weighting adjustments are used for all other sample persons with missing waves (Singh *et al.* 1990). Further discussion of methods of handling wave nonresponse in panel files is provided by Lepkowski (1989), Kalton (1986), and Lepkowski *et al.* (1993).

Another complication of some household panel surveys is the occurrence of partial household nonresponse, which occurs when the survey data are collected for some but not all members of a sampled household at a particular wave.

The lack of data for one individual in a household means that key household characteristics (e.g., household earnings) cannot be computed. One solution is to drop the household and its responding members from the sample, and use a weighting adjustment. Another is to impute the responses for the nonresponding household members, as is done in SIPP (where they are termed Type z nonrespondents). With the latter solution, data are available for all members of responding households, and hence person-level adjustments are unnecessary within responding households.

We now turn to consider the issues involved in dealing with missing data for cross-sectional analyses of a household panel survey. A separate cross-sectional file containing data for all responding households and their members (either deleting the households or imputing values for missing responses in the case of partial household nonresponse) can be created for each wave. Adjustments are then needed to compensate for the nonresponding and uncovered households and persons in each file.

Nonresponding households at wave t can be divided into total nonrespondents and wave nonrespondents. Total nonresponse occurs in a panel survey when a sampled element fails to provide data for any wave. Since it is common practice not to follow up sample households that fail to respond at the initial wave, these households and their members are generally the total nonrespondents. Compensation for total nonrespondents is relatively straightforward. The Wave 1 weights of the responding households at the initial wave can be adjusted using standard nonresponse adjustment methods and the adjusted weights can be used instead of the selection probabilities in developing the cross-sectional weights for later waves. Most nonresponse adjustment methods, such as weighting class adjustments (Kalton and Kasprzyk 1986) and adjustments based on response propensities (Little 1986), are based on the assumption that nonresponse is random within weighting classes or that the probabilities of responding within a class can be estimated precisely. Under these conditions, the response mechanism can be treated as an additional stage of sampling. Thus, the selection probabilities, p_j , used to define the weights in equation (3.5) may be redefined as the product of the selection probabilities and the adjustment due to nonresponse. For example, if weighting class adjustments are used, the selection probability of original household h_j multiplied by the weighted response rate for the weighting class in which h_j falls is used instead of the original p_j . The previous results then follow for the weights adjusted for total nonresponse.

The same approach can also be extended to cover weighting adjustments for households responding at the initial wave that lead to no responding households at wave t . In this case, the responding households at the initial wave can be divided into weighting classes based on responses given at that wave, and the weights of households leading to one or more responding households at wave t

can be further adjusted to compensate for those leading to no responding households at wave t . The revised w'_{ij} can then be employed in equation (3.5) and subsequently.

Both the above nonresponse adjustments are applied in relation to the original households. A further type of household nonresponse cannot be handled in this way. This type of nonresponse involves the situation where an original household splits into two or more separate households at wave t , and where some but not all of those households respond at that wave. In this case the adjustment for the nonresponding households needs to be made in relation to the wave t households, H_t , rather than the original households, h_j . If the number of original households having members in each wave t nonresponding household of this type were known, the weights w_i for these households could be computed using the approach described above. Then weighting adjustments could be readily applied within weighting classes of the wave t households to compensate for the nonresponding households. In practice, however, the number of original households having members in a nonresponding household at wave t may often be unknown. One approach for handling this situation is to estimate this number by the average number for responding households at wave t that have similar characteristics to (e.g., they are also splits from original households), and are in the same weighting class as, the nonresponding household. Using such estimated numbers where necessary, the weights w_i can be determined for all nonresponding households of the type being discussed. Standard weighting adjustments can then be applied to the responding households at wave t to compensate for these nonresponding households.

Incomplete coverage of the target population is another nonsampling problem that has been traditionally addressed in surveys by adjusting the sampling weights. For example, poststratification (see, for example, Holt and Smith 1979) and generalized raking procedures (Deville *et al.* 1993) are often used to adjust the weights so that they sum to counts from independent sources not subject to undercoverage. These adjustments may also reduce the sampling errors of the estimates, although bias reduction is often more critical.

The control totals used in most household surveys are counts of the number of persons in classes defined by characteristics such as age, sex and race. This method of reducing undercoverage bias may be fully sufficient when estimates of persons are the only types of statistics to be produced from the survey. However, further steps are needed to calculate household-level weights for producing statistics of household characteristics.

One approach to developing household-level weights when control totals are based on person-level counts is called the principal person method, as described by Alexander (1987). In this method, poststratification adjustments are applied at the person level. One household member is then identified as the principal person and the fully adjusted

weight for that person is assigned to be the household weight. Since the person weights are already adjusted to the control totals, the household weight does incorporate some adjustments to reduce coverage bias. For cross-sectional estimation from a household panel survey, the principal person method can readily be used in conjunction with the equal household and person weighting schemes to produce household level weights.

A disadvantage of the principal person method is that estimates of the number of persons calculated using the principal person weight will generally differ from control totals. The estimates may also differ significantly depending on the criteria used to identify the principal person in the household. The method has also been criticized because the weights of the members of the same household differ, even though they were all selected at the same rate as the household. Estimation schemes proposed by Alexander (1987), Lemaître and Dufour (1987), and Zieschang (1990) address these objections by constraining the household weights so that they are consistent with the independent person-level totals while minimizing the distance between the original household weights and the adjusted weights. All three consider variants of a generalized least squares (GLS) algorithm to achieve this objective. Zieschang (1990) shows how GLS can be used to create weights that are consistent with the person controls and force all persons within a household to have the same weight.

The application of GLS methods when the household weights are computed using the equal household and person weighting schemes is relatively straightforward. However, empirical evaluation of the consequences of using these methods is needed. The GLS methods have the unattractive feature that they can result in negative weights. Furthermore, the increase in the variation in the weights arising from the constraints imposed may result in less precise estimates. This concern may be especially important when the variability in the household weights is increased due to their multiple routes for selection and the equal household or person weighting schemes are necessary.

7. SUMMARY AND CONCLUDING REMARKS

This paper has described weighting schemes for cross-sectional analysis of later waves of a household panel survey using data for all households for which and all individuals for whom data are collected. These weighting schemes can accommodate new entrants to the population who move in to live with members of the original population, but not other new entrants.

The usual inverse selection probability weighting scheme requires information on the household selection probabilities of all members of the households sampled at a later wave, as well as the joint selection probabilities of the original households that contribute members to the later

wave households. The inverse selection probability weighting scheme can often not be applied because these probabilities are unknown. To deal with this problem, an alternative approach that requires information on only the selection probabilities of sampled original households is described.

This alternative approach produces a class of weighting schemes including the equal person (fair share) scheme used in SIPP and the equal household weighting scheme. All the schemes in this class produce weights that are in expectation equal to those produced by the usual inverse selection probability scheme. The variance in the weights around the inverse selection probability weights gives rise to an increase in the variance of the survey estimates. When the original households are selected with approximately equal probability, the equal household weighting scheme is near optimal for both household and individual level analyses to control this increase in variance.

The alternative class of weighting schemes produces unbiased estimates of population totals for any choice of constant α_{ij} that satisfies the condition $\sum_j \alpha_{ij} = 1$ and for any initial sample design. The equal household and equal person weighting schemes are, however, suboptimal for non-epsem initial samples. One of them may nevertheless be the appropriate scheme for such designs, because the optimal choice of the α_{ij} depends on the unknown initial selection probabilities, and hence cannot be determined. The equal household and equal person weighting schemes have different data requirements, in that the former requires knowledge of the number of Wave 1 households represented in the Wave t household whereas the latter does not. The fact that this information may not always be readily obtainable thus argues in favor of the equal person weighting scheme.

The cross-sectional individual weights for a particular wave can be used as the starting weights for a longitudinal analysis that begins at that wave. This procedure includes cohabitants present at that wave in the longitudinal analysis. However, if cohabitants are not followed when they cease to live with sampled persons, those who leave sample persons before the end of the period of the longitudinal analysis become nonrespondents. Before cohabitants are included in a longitudinal analysis, a check should therefore be made to ensure that their inclusion will not give rise to risks of serious nonresponse bias.

The class of weighting schemes described has a broader range of application than that indicated here. It can in fact be usefully applied in any situation where an inverse selection probability weighting scheme would be appropriate, but where not all the inclusion probabilities and joint inclusion probabilities are known. Consider, for instance, the modified version of the Mitofsky-Waksberg random digit dialing sampling procedure for telephone surveys described by Brick and Waksberg (1991). A sample of telephone numbers (primes) is selected at the first stage of this two-stage sample design. If a prime number is found

to be a working residential number, that household is selected and a fixed number of additional telephone numbers in the same 100-bank is selected. The households found at these numbers are then all included in the sample. If a prime number is not a working number, the sampling process stops. With this procedure, the probability of a working residential number being selected depends on the number of working residential numbers in its 100-bank, and hence differs across 100-banks. This probability can be estimated from the sample of telephone numbers in the 100-bank. A complication arises, however, when a sampled household has two or more telephone numbers. In this case, the selection probability of the sampled telephone number can be estimated, but those of the nonsampled numbers cannot. Thus, the standard inverse selection probability weighting scheme cannot be used. However, the alternative weighting scheme described here can be employed.

ACKNOWLEDGEMENTS

We thank the referees for their helpful comments on the earlier version of the paper. The paper reports research undertaken for the U.S. Census Bureau. The views expressed are the authors'. They do not necessarily reflect the views of the Census Bureau.

REFERENCES

- ALEXANDER, C.H. (1987). A class of methods for using person controls in household weighting. *Survey Methodology*, 13, 183-198.
- BRICK, J.M., and WAKSBERG, J. (1991). Avoiding sequential sampling with random digit dialing. *Survey Methodology*, 17, 27-41.
- BUCK, N., GERSHUNY, J., ROSE, D., and SCOTT, J. (Eds.) (1994). *Changing Households: The British Household Panel Survey 1990-1992*. Colchester, U.K.: ESRC Research Centre on Micro-social Change.
- CITRO, C.F., and KALTON, G. (1993). *The Future of the Survey of Income and Program Participation*. Washington D.C.: National Academy Press.
- DEVILLE, J.-C., SÄRNDAL, C.-E., and SAUTORY, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.
- DUNCAN, G.J., and HILL, M.S. (1985). Conceptions of longitudinal households: Fertile or futile? *Journal of Economic and Social Measurement*, 13, 361-375.
- ERNST, L.R. (1989). Weighting issues for longitudinal household and family estimates. In *Panel Surveys*, (Eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh). New York: John Wiley, 139-159.

- GAILLY, B., and LAVALLÉE, P. (1993). *Insérer des Nouveaux Membres dans un Panel Longitudinal de Ménages et D'Individus: Simulations*. Walferdange, Luxembourg: CEPS/Instead.
- HILL, M.S. (1992). *The Panel Study of Income Dynamics: A User's Guide*. Newbury Park, CA: Sage Publications.
- HILL, M.S. (1995). Personal Communication.
- HOLT, D., and SMITH, T.M.F. (1979). Post stratification. *Journal of the Royal Statistical Society, A*, 142, 33-46.
- HUANG, H. (1984). Obtaining cross-sectional estimates from a longitudinal survey: Experiences of the Income Survey Development Program. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 670-675.
- JABINE, T.B., KING, K.E., and PETRONI, R.J. (1990). *Survey of Income and Program Participation: Quality Profile*. Washington D.C.: U.S. Bureau of the Census.
- JUDKINS, D., HUBBLE, D., DORSCH, J., MCMILLEN, D., and ERNST, L. (1984). Weighting of persons for SIPP longitudinal tabulations. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 676-687.
- KALTON, G. (1986). Handling wave nonresponse in panel surveys. *Journal of Official Statistics*, 2, 303-314.
- KALTON, G., and KASPRZYK, D. (1986). The treatment of missing survey data. *Survey Methodology*, 12, 1-16.
- KASPRZYK, D. (1988). *The Survey of Income and Program Participation: An Overview and Discussion of Research Issues*. SIPP Working Paper No. 8830. Washington D.C.: U.S. Bureau of the Census.
- KASPRZYK, D., and MCMILLEN, D.B. (1987). SIPP: Characteristics of the 1984 Panel. *Proceedings of the Social Statistics Section, American Statistical Association*, 181-186.
- KISH, L. (1965). *Survey Sampling*. New York: Wiley.
- LAVALLÉE, P. (1995). Cross-sectional weighting of longitudinal surveys of individuals and households using the weight share method. *Survey Methodology*, 21, 25-32.
- LAVALLÉE, P., and HUNTER, L. (1992). Weighting for the Survey of Labour and Income Dynamics. *Proceedings: Symposium 92, Design and Analysis of Longitudinal Surveys*, Statistics Canada, 65-75.
- LAVALLÉE, P., MICHAUD, S., and WEBBER, M. (1993). The Survey of Labour and Income Dynamics, design issues for a new longitudinal survey in Canada. *Bulletin of the International Statistical Institute*, 49th Session, Contributed Papers, Book 2, 99-100.
- LEMAÎTRE, G., and DUFOUR, J. (1987). An integrated method for weighting persons and families. *Survey Methodology*, 13, 199-207.
- LEPKOWSKI, J. (1989). Treatment of wave nonresponse in panel surveys. In *Panel Surveys*, (Eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh). New York: Wiley, 348-374.
- LEPKOWSKI, J.M., MILLER, D.P., KALTON G., and SINGH, R. (1993). Imputation for wave nonresponse in the SIPP. *Proceedings: Symposium 92, Design and Analysis of Longitudinal Surveys*, Statistics Canada, 99-109.
- LITTLE, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review*, 54, 139-157.
- LITTLE, R.J.A. (1989). Sampling weights in the PSID: Issues and comments. Panel Study of Income Dynamics Working Paper, Ann Arbor: University of Michigan.
- NELSON, D., MCMILLEN, D., and KASPRZYK, D. (1985). *An Overview of the SIPP, Update 1*. SIPP Working Paper No. 8401. Washington D.C.: U.S. Bureau of the Census.
- SINGH, R., HUGGINS, V., and KASPRZYK, D. (1990). *Handling Single Wave Nonresponse in Panel Surveys*. SIPP Working Paper No. 9009. Washington D.C.: U.S. Bureau of the Census.
- ZIESCHANG, K.D. (1990). Sample weighting methods and estimation of totals in the Consumer Expenditure Survey. *Journal of the American Statistical Association*, 85, 986-1001.

Modelling Net Undercoverage in the 1991 Canadian Census

PETER DICK¹

ABSTRACT

In 1991, Statistics Canada for the first time adjusted the Population Estimates Program for undercoverage in the 1991 Census. The Census coverage studies provided reliable estimates of undercoverage at the provincial level and for national estimates of large age – sex domains. However, the population series required estimates of undercoverage for age – sex domains within each province and territory. Since the direct survey estimates for some of these small domains had large standard errors due to the small sample size in the domain, small area modelling techniques were needed. In order to incorporate the varying degrees of reliability of the direct survey estimates, a regression model utilizing an Empirical Bayes methodology was used to estimate the undercoverage in small domains. A raking ratio procedure was then applied to the undercoverage estimates to preserve consistency with the marginal direct survey estimates. The results of this modelling process are shown along with the estimated reduction in standard errors.

KEY WORDS: Small area; Empirical Bayes; Undercoverage.

1. INTRODUCTION AND BACKGROUND

The Census of Canada is conducted every five years; one of its objectives is to provide the Population Estimates Program with accurate baseline counts of the number of persons by age and sex within each province and territory. Unfortunately, not all eligible persons are correctly enumerated by the Census. As part of the evaluation of the Census, Statistics Canada estimates, through two sample surveys, the net number of persons missed by the Census. The estimates are from the Reverse Record Check Study, which estimates the gross number of persons missed by the Census, and the Overcoverage Study, which estimates persons double counted or erroneously included in the final Census count. When combined the figures estimate the net number of people missed by the Census.

These surveys were designed to produce reliable direct estimates for large areas, such as provinces, and for large domains, such as age – sex combinations at the national level. However, the Population Estimates Program requires estimates of missed persons for single year of age for both sexes for each province. However using the direct survey estimate would result in individual estimates having unacceptably high standard errors due to insufficient sample in the small domain. One approach to reducing the variance of the small domain estimates would be to borrow strength from related domains. This approach leads to creating an explicit model for the small domain that can be used to predict the net missed persons in that domain.

The result of modelling the small domain estimates is to produce a series of estimates with a smaller Mean Square Error than the direct estimate. However, as opposed to the

direct survey estimate which is design unbiased, the modelling approach will introduce a bias for each estimate. Thus modelling the small domain estimates implies that a trade off is required between reducing the variance of each estimate and the bias introduced through the modelling process. One approach to ensuring that the more reliable direct survey estimates are utilized is to introduce an Empirical Bayes model. This procedure creates an estimate that is a combination of a model estimate and the direct survey estimate weighted by their respective variances. It is an Empirical Bayes estimate instead of a Bayes estimate because underlying parameters are first estimated, then these estimated parameters are considered known in later calculations. Note that since the individual sampling variances are used in the estimation, a more precise direct estimate would contribute much more to the final Empirical Bayes estimate than a similar estimate with low precision. This ensures that the model does not dominate estimates that are already considered reliable. It is also possible to approach this estimation problem through a Hierarchical Bayes methodology: details on this method can be found in Datta, *et al.* (1992). Ghosh and Rao (1994) give an appraisal of both the Hierarchical Bayes and Empirical Bayes approaches to small area estimation.

Outside of Canada, two different approaches to smoothing the Census undercoverage have been described in the literature. In the United States, the net undercoverage in the 1990 American Census was evaluated by means of the Post Enumeration Survey (Hogan 1992). Initially, it was planned to multiply the US Census counts by adjustment factors (the ratio of true population over the enumerated population) for 1,392 *a priori* defined post strata.

¹ Peter Dick, Social Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6.

These estimated adjustment factors would then be used to adjust the Census count for missed persons. Since some of these 1,392 estimated adjustment factors had high standard errors, it was proposed to smooth the direct estimates through an Empirical Bayes regression model, similar to one proposed by Ericksen and Kadane (1985), and then to rake the smoothed estimates to agree with direct estimates for large geographic regions. However, this approach was criticized by Freedman and Navidi (1992). Eventually, the United States Department of Commerce, the U.S. Census Bureau's parent agency, decided not to proceed with adjusting the Census counts for underenumeration in July 1991. Consideration was also given in the United States to adjusting the post Censal population estimates for undercoverage in the Census, but the Department of Commerce also rejected this adjustment.

The Australians use a different method than the Americans for estimating the domain totals. Choi, Steel and Skinner (1988) describe a methodology that incorporates the estimates of net undercoverage of the Census into the population estimates but leaves the actual Census counts as enumerated. The under enumeration is estimated through a Post Enumeration Survey (PES) and demographic analysis. The small domain estimates are produced by raking the Census age counts for each sex to the PES estimates for national age/sex totals and part of State/Territory/sex totals.

The procedure proposed for the 1991 Canadian Census combines some of the elements of both the American and Australian approaches. As in the American procedure, a model is postulated for the underlying true adjustment factors and another model is postulated for relating the direct survey estimates to the true underlying adjustment factors. Through Empirical Bayes estimation, a new smoothed adjustment factor is estimated that will have a lower MSE than the direct survey estimate. These smoothed adjustment factors are then converted into estimates of missed persons. The Australian method for constraining the resulting estimates to known marginal totals is then adopted. These final raked estimates are used as the base for the small domain estimates of missed persons. In turn, these estimates are adjusted to account for known demographic principles (See Michalowski 1993). Details on the technical criteria for adjustment of the population estimates can be found in Royce (1992).

This paper is organized as follows. In Section 2, some background information on the two sample surveys is described and the basic Empirical Bayes model is presented. Assumptions and limitations of the model are also discussed and the estimation of the parameters is briefly discussed. In Section 3, the explanatory variables used in the regression model are presented and the model building process is described. The final model is presented and the results displayed. Section 4 presents a discussion on the rationale behind constraining the Empirical Bayes estimates to

reliable marginal totals. The final adjusted estimates are then presented. Finally, Section 5 presents some conclusions and topics for further study.

2. MODEL FOR THE ADJUSTMENT FACTORS

2.1 Background and Notation

The model for the adjustment factors requires input data. The actual data originates with two coverage studies: the Reverse Record Check (RRC) and the Overcoverage Study (OCS). The RRC is used to estimate the number of persons missed by the Census while the OCS is used to estimate the number of persons erroneously included in the Census count. These surveys are designed to give reliable estimates of net undercoverage for all provinces, some of the larger metropolitan areas and for some large national domains, such as males aged 20 to 24. Since the surveys are independent, it can be assumed that the variance of net missed persons will be the sum of the two estimated variances from the RRC and the OCS. Further details on these studies can be found in Germain and Julien (1993) and the 1991 Census Technical Report – Coverage (Statistics Canada 1994).

The domains of interest can be defined by partitioning the sample into $p = 1, 2, \dots, P$ provinces/territories and $a = 1, 2, \dots, A$ age – sex groups, hence a total of $A \times P$ domains require estimates. Let C_i be the number of persons in the i -th province – age domain enumerated in the Census and T_i be the true population of the same domain. The net number of persons missed in the i -th cell is $M_i = T_i - C_i$. The adjustment factor, Θ_i , is the ratio of the true population in a domain over the Census count, while the undercoverage rate, U_i , the unit that is usually reported in the releases from the coverage studies, is the ratio of missed persons over the true population.

The true adjustment factors, Θ_i , which are the variables that we wish to estimate, can be written as:

$$\Theta_i = \frac{T_i}{C_i} = \frac{M_i + C_i}{C_i}.$$

Undercoverage rates (U_i) which are usually reported in the releases from Statistics Canada, are related to the adjustment factors through the relationship

$$U_i = M_i(M_i + C_i)^{-1} = 1 - \Theta_i^{-1}.$$

In the modelling of the adjustment factors, the creation of ultimate domains is required. These domains are those at which the actual direct survey estimates of the adjustment factors will be produced. There must be an estimate for each province (10) and territory (2), so immediately P is fixed at 12. The age groups were fixed at 4 to create

national estimates that have acceptably low standard errors. These age groups are defined for male and female as follows: 0 to 19 years of age; 20 to 29 years of age; 30 to 44 years of age; and 45 years and older. In total there are $12 \times 8 = 96$ direct survey estimates of adjustment factors that have to be fitted into the Empirical Bayes model. Each domain requires, besides the direct estimate of the adjustment factor, an associated estimate of the sampling variance.

2.2 Model and Assumptions

The basic model for the undercount is composed of two distinct parts. The first part describes how the direct survey estimates are related to the true underlying adjustment factors, while the second part models the relationship between the true adjustment factors and a set of explanatory variables. Since the parameters in the regression model are estimated by first estimating the parameters of an assumed underlying prior distribution and then assuming that these estimated parameters are known for any further calculation, this model is known as an Empirical Bayes model (Maritz and Lwin 1989).

The first part of the model, the sampling model, relates the observed adjustment factors to the true adjustment factors. This relationship is assumed true within each domain, and can be expressed as:

$$\begin{aligned} &\text{the observed adjustment factor} = \\ &\text{the true adjustment factor} + \text{a random error.} \end{aligned}$$

The sampling model is written as follows:

$$F_i = \Theta_i + \epsilon_i : \epsilon_i \sim \text{Normal}(0, \sigma_i^2),$$

$$i = 1, 2, \dots, n = A \times P,$$

where Θ_i is the true adjustment factor and ϵ_i is a random error component with a variance of σ_i^2 . The assumptions underlying this model are:

- (a) the sampling errors, ϵ_i , have mean zero;
- (b) the sampling variances, σ_i^2 , are known in each of the n domains;
- (c) since the sample was selected independently within each domain, the covariance between the sampling errors ϵ_i in domain i and ϵ_j in domain j is zero; and
- (d) the random errors ϵ_i are normally distributed in each domain.

Further discussion on the assumption of the known sampling variance in each domain is given below.

The second part of the model, the regression model, relates the true adjustment factors to a set of underlying explanatory variables. This model states that:

$$\begin{aligned} &\text{the true adjustment factor} = \text{a linear combination} \\ &\text{of explanatory variables} + \text{a random error.} \end{aligned}$$

The regression model can be written as:

$$\begin{aligned} \Theta_i &= X_i \beta + \delta_i : \delta_i \sim \text{Normal}(0, \tau^2), \\ i &= 1, 2, \dots, n = A \times P, \end{aligned}$$

where X_i is the i -th row in X , a known $(n \times p)$ matrix of explanatory variables, β is a $(p \times 1)$ vector of unknown regression parameters and δ_i is (a different) random error with a model variance of τ^2 . Underlying the system model are the following assumptions:

- (a) the model errors, δ_i , have mean zero;
- (b) the model variance, τ^2 , is constant over all n domains;
- (c) the model errors, δ_i , are normally distributed;
- (d) the model errors, δ_i , are independent of sampling errors, ϵ_i ;
- (e) the covariance between different domains is zero (*i.e.*, $\text{Cov}(\delta_i, \delta_j) = 0$).

The problem is to use both the sampling model and the regression model to estimate Θ_i , the true adjustment factors. The conditional expectation for Θ_i given β , σ_i^2 , τ^2 , F_i can be determined for the joint model. Using standard arguments (Rao 1973), it can be shown that the conditional expectation of Θ_i is:

$$E(\Theta_i | \beta, \sigma_i^2, \tau^2, F_i) = (1 - \omega_i)X_i \beta + \omega_i F_i, \quad (1)$$

where $\omega_i = \tau^2(\tau^2 + \sigma_i^2)^{-1}$.

Equation (1) is the basis for all the estimates that follow, although a few modifications need to be made before applying it to the data. Note that it is basically a weighted average of the direct survey estimate and the regression model estimate of the adjustment factor. Each estimate is weighted according to the precision with which it was estimated. If the sampling error, σ_i^2 , is small compared to the model error, τ^2 , implying that the direct survey estimate is relatively precise, then the final smoothed estimate will be mainly composed of the direct survey estimate. However, if the direct survey estimate has a large sampling variance relative to the model variance then the final smoothed estimate will be mainly constituted from the best linear unbiased predictor. The amount each estimate contributes to the final smoothed estimate is controlled by the weighting coefficient, ω_i .

Some limitations apply to interpretations that can be made about this model. First, it must be emphasized that this model is purely descriptive; it cannot be considered to be a causal model. Since the primary goal of this model is descriptive, the inferences on the regression parameters, β , while interesting are not of primary importance. Hence, the final regression model when it contains a term, say, on British Columbia renters and not Manitoba renters, is only saying that British Columbia renters explain a

significant portion of the variation in adjustment factors in British Columbia while Manitoba renters does not explain a significant portion of the variation in adjustment factors in Manitoba.

As mentioned above, the sampling variances associated with the direct survey estimates of the adjustment factors are considered known in the Empirical Bayes model. However, experience has shown that the directly estimated variances are, in fact, somewhat unstable. In order to create some stability with the estimation of these variances it is proposed to model them. If we consider the design of the two sample surveys, then, under relatively mild assumptions, Dick (1993) has shown that within each domain the variance of the estimate of missed persons is proportional to a power of the Census count. If we add in appropriate normalizing parameters, then this relationship can be written as:

$$\sigma_i^2 C_i^2 = V(M_i) = K C_i^\gamma,$$

or, as in the form of a regression equation,

$$\begin{aligned} \text{Log}(V(M_i)) &= \alpha + \gamma \log(C_i) + \eta_i \quad \text{with} \\ \eta_i &\sim N(0, \zeta^2). \end{aligned}$$

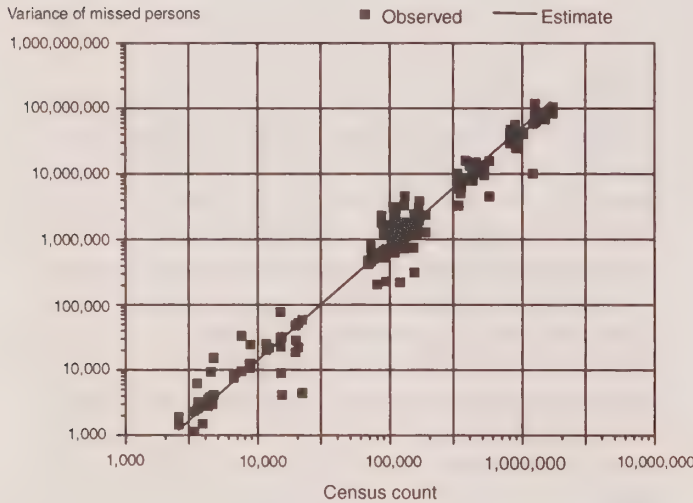


Figure 1. Observed variance vs. census.

This model for the sampling variance assumes that the product of the design effect and the undercoverage rate is constant within each domain. As discussed in Dick (1993), this assumption appears to be reasonable. Figure 1 shows the plot of the observed variance of missed persons calculated from the two coverage studies versus the Census count for the 96 domains. The least squares regression line was estimated as

$$\log(\hat{v}(M_i)) = -6.133 + 1.715 \log C_i$$

and is also plotted in Figure 1. A residual analysis (Dick 1993) did not detect any apparent violations of the underlying model assumptions. Since, in addition, the coefficient of determination, the R^2 , is 0.943, this model was adopted for producing the sampling variances. The estimated survey variances were calculated for the adjustment factors through

$$\hat{v}(F_i) = \hat{v}(M_i) / C_i^2.$$

It will be assumed that these predicted values for the sampling variances are the actual 'known variances' required for the Empirical Bayes model.

2.3 Parameter Estimation

So far the model has been described in purely Bayesian terms: only the parameter Θ_i is considered unknown. Taking the usual Empirical Bayes approach (Maritz and Lwin 1989), we will assume that all the parameters except β , the regression parameter, are known. The conditional expectation of Θ_i with the regression parameter estimated can be written as

$$\tilde{F}_i^{(eb)} = E(\Theta_i \mid \hat{\beta}, \sigma_i^2, \tau^2, F_i).$$

However, in practice, the model variance, τ^2 , is also unknown and must be estimated. The conditional expectation of Θ_i will now change to

$$\hat{F}_i^{(eb)} = E(\Theta_i \mid \hat{\beta}, \sigma_i^2, \hat{\tau}^2, F_i),$$

where the sampling variance, σ_i^2 , is still considered known.

To estimate the model variance and the regression coefficients in the Empirical Bayes model, the marginal distribution of the observed adjustment factors, $m(F_i) \sim N(x_i \beta, \tau^2 + \sigma_i^2)$, can be used. Three possible methods were examined for estimating the variance parameter, τ^2 : Method of Moments (MM) as in Fay and Herriot (1979), Maximum Likelihood (MLE) as in the PES in the United States (Hogan 1992) and Restricted Maximum Likelihood (REML).

It is well known that MLE estimation of variance components is biased downwards (Harville 1977). Underestimation of the model variance in the Empirical Bayes model would result in more reliance being placed upon the regression model instead of the direct survey estimate. This is a result we wished to avoid. In Dick (1993), it is shown that there is little difference between the estimates of the model variance from REML or MM. Since the REML has a well understood asymptotic theory, it was adopted for the estimation of the model variance in the Empirical Bayes model.

Harville gives a full account of REML estimation. The basic approach is to first estimate the regression parameter, and then to estimate the model variance from the resulting residuals instead of the actual data. If we let X^* be a matrix

of $(n - p)$ linear contrasts such that $E[X^{*'}F] = 0$, then Harville shows that the resulting (log) likelihood function, L_{reml} , when maximized with respect to the unknown model variance will give the restricted maximum likelihood estimates.

In the context of the Empirical Bayes model, Harville's approach can be described as follows. First, an initial estimate, usually zero, of the model variance, $\hat{\tau}_{(0)}^2$, is made and then the regression parameter, β , is estimated through weighted least squares:

$$\hat{\beta}_{(1)} = (X^t \hat{V}_0^{-1} X)^{-1} X^t \hat{V}_0^{-1} F, \quad (2)$$

where $\hat{V}_0 = \text{diag}(\hat{\tau}_{(0)}^2 + \sigma_i^2; i = 1, \dots, n)$. Using this estimate of $\hat{\beta}_{(1)}$, a new REML estimate of the model variance, $\hat{\tau}_{(1)}^2$, can be made through

$$\hat{\tau}_{k+1}^2 = \hat{\tau}_k^2 + \left(\frac{\partial L_{\text{reml}}}{\partial \tau^2} \right) [i(\tau^2)]^{-1}, \quad k = 0, 1, \dots, \quad (3)$$

where, if we set $\hat{P}_k = \hat{V}_k^{-1} - \hat{V}_k^{-1} X (X^t \hat{V}_k^{-1} X)^{-1} X^t \hat{V}_k^{-1}$, we have

$$\frac{\partial L_{\text{reml}}}{\partial \tau^2} = -\frac{1}{2} \text{trace } \hat{P}_k + \frac{1}{2} (F - X\hat{\beta})^t \hat{V}_k^{-1} \hat{V}_k^{-1} (F - X\hat{\beta})$$

and

$$i(\tau^2) = -E \left[\frac{\partial^2 L_{\text{reml}}}{(\partial \tau^2)^2} \right] = \frac{1}{2} \text{trace } (\hat{P}_k^t \hat{P}_k).$$

Note, upon convergence of τ^2 and β , $i(\tau^2)^{-1}$ will be the asymptotic variance of $\hat{\tau}^2$.

By iterating between (2) and (3), new estimates of τ^2 will be used to update the estimate of β , which in turn will be used to update the estimate of τ^2 . The iterations then continue until a suitable convergence has been reached: in this case $((\hat{\tau}_{k+1}^2 / \hat{\tau}_k^2) - 1) < 10^{-6}$ was used.

Once the estimates for $\hat{\beta}$, the regression parameters, and $\hat{\tau}^2$, the model variance, have been determined, then the final smoothed estimates can be found. Maritz and Lwin (1989) show that the Empirical Bayes, or smoothed, estimate can be written as

$$\hat{F}_i^{\text{eb}} = (1 - \hat{\omega}_i) X_i \hat{\beta} + \hat{\omega}_i F_i,$$

where $\hat{\omega}_i = \hat{\tau}^2 (\hat{\tau}^2 + \sigma_i^2)^{-1}$. This is a combination of the original estimate and the regression estimate weighted by their respective variances.

The objective of the smoothing model is to create a series of estimates with smaller MSE than the original estimates. Prasad and Rao (1990), through asymptotic arguments, have suggested using the following estimator for the mean square error:

$$\text{MSE}[\hat{F}_i^{\text{eb}}] = \text{MSE}[\tilde{F}_i^{\text{eb}}] + \left[\left(\frac{\partial \omega_i}{\partial \tau^2} \right)^2 \omega_i E(\hat{\tau}^2 - \tau^2)^2 \right].$$

The mean square error for the Empirical Bayes estimate, using restricted maximum likelihood estimation, has been conjectured by Cressie (1992) to be:

$$\widehat{\text{MSE}}[\hat{F}_i^{\text{eb}}] = \widehat{\text{MSE}}(\tilde{F}_i^{\text{eb}}) + 2 \hat{g}_{3i}(\hat{\tau}^2) = \hat{g}_{1i}(\hat{\tau}^2) + \hat{g}_{2i}(\hat{\tau}^2) + 2 \hat{g}_{3i}(\hat{\tau}^2),$$

where

$$\hat{g}_{1i}(\hat{\tau}^2) = \hat{\tau}^2 (1 - \hat{\omega}_i)$$

$$\hat{g}_{2i}(\hat{\tau}^2) = (1 - \hat{\omega}_i)^2 X_i^t (X^t \hat{V}^{-1} X)^{-1} X_i$$

and

$$\hat{g}_{3i}(\hat{\tau}^2) = (1 - \hat{\omega}_i)^2 (\hat{\tau}^2 + \sigma_i^2)^{-1} [i(\tau^2)]^{-1}.$$

The assumed normality of ϵ_i and δ_i is an important assumption in the derivation. Note the value for the sampling variance, σ_i^2 , is assumed known.

Prasad and Rao give the following interpretation to each of the three components: $\hat{g}_{1i}(\hat{\tau}^2)$ is the Bayes estimate of the variance, $\hat{g}_{2i}(\hat{\tau}^2)$ is the contribution from estimating the regression parameters and $\hat{g}_{3i}(\hat{\tau}^2)$ is the contribution from estimating the model variance τ^2 . An estimate of the component due to the estimation of the sampling variance is *not* available: the additional variance this would add is not known but its absence clearly implies that the MSE is underestimated.

3. EMPIRICAL BAYES LINEAR MODEL

3.1 Explanatory Variables

The Empirical Bayes model described above was fitted to the 96 observed adjustment factors, with the sampling variances estimated using the method described in Section 2. The linear model that was fitted to this data included the following explanatory variables:

- An indicator variable for each province/territory.
- An indicator variable for each sex.
- An indicator variable for each age group.
- A variable indicating the percentage of people in the domain that are renters.
- A variable indicating the percentage of people in the domain that do not speak either official language.
- Various interaction variables including province by renters, province by non-official language, age and sex by renters.

In total, 42 variables were used in the initial regression.

These variables were selected for the initial regression model based, in part, on the experiences of previous RRC studies (Burgess 1988), partly on the results of the 1991 coverage studies (Germain and Julien 1993) and partly on the experiences of the PES in the United States as described in Hogan (1992) and Datta *et al.* (1992). The actual rationale for the variables to be included are as follows:

- (a) The province indicator was included as an indication of the difficulty of Census collection within each province. Prior to the 1991 Census, it was assumed that collection would be more difficult in British Columbia and Ontario, and the anecdotal field evidence during collection seemed to support this conjecture.
- (b) The age and sex variable were included because of the known differences in undercoverage rates between males and females. The undercoverage, in previous studies, has also shown a marked increase for individuals in their 20's.
- (c) Tenure, in effect the percent of renters in each domain, was included because of the experiences in the United States PES, results of previous RRC studies and as a suggestion from the Statistics Canada Statistical Methods Advisory Committee.
- (d) The use of non-official language was an attempt to locate the immigrant and minority groups that in the past have tended to have higher undercoverage rates.
- (e) The interaction terms were included to further refine the predictive power of the model.

The mean encompasses all those variables that are not included in the model. Note that since indicator variables are used for province, sex and age-sex, one variable has to be excluded in order to avoid a singular design matrix. In effect, the missing variable, say the province indicator for Newfoundland, is included in the mean.

An operational constraint was also placed on the model. The SAS IML program written to estimate the parameters was limited to 4,095 numeric elements in the design matrix, hence with 96 domains, or observations, the model was limited to a maximum of 42 variables.

3.2 Model Building Process

After starting with the full regression model and 42 explanatory variables, a procedure was needed to remove those variables that were not statistically significant. The procedure chosen was to eliminate the least significant variable after each completed estimation cycle. This implies that for the 42 variable model, the variable Female Renters aged 0 to 19 would be eliminated since it has a t -value of 0.05. The regression model was then re-run with the remaining 41 variables. The least significant variable was then eliminated from that model. This procedure is equivalent to the Backward Stepwise Regression described in Draper and Smith (1966, page 167).

The Backward Stepwise Regression method was used to eliminate all variables until all remaining variables had a t -values greater than 2 (in absolute value). However when the final model was examined, it was noticed that a multicollinearity problem existed between the indicator variables for certain provinces and the interaction terms for renters within the same provinces. The implication of this problem is that there are some explanatory variables which are highly correlated with each other. This in turn implies that not all parameters in the model can be estimated precisely. As a rule of thumb Judge *et al.* (1984, page 459) suggest that this can be a problem when the simple correlation between variables is greater than R^2 , the coefficient of determination. The final model had a $R^2 = 0.85$ and the simple correlation between the variables in question were all greater than 0.90 (in absolute value, since the correlations were negative).

A solution to this problem was to delete the variables with the lower t -values which turned out to be the provincial indicators. The final model is shown in Table 1 with the estimated coefficients and their t -values. The effect of removing the provincial indicators was to lower the final R^2 from 0.85 to 0.844, thus little predictive power has been lost.

Table 1
Final Estimates of Variables Used in Regression

Category	Variable	Final Estimate ($\hat{\beta}$)	T-Value (absolute value) ($H_0: \beta = 0$)
Mean	Mean	1.0075	575.72
Age - Sex Combination	Male 20 to 29	0.0563	15.34
	Male 30 to 44	0.0208	5.81
	Female 20 to 29	0.0240	6.49
Sex by Age by Non-Official Language	Female Language 0 to 19	0.0797	2.75
Tenure by Province	British Columbia Renters	0.0449	3.96
	Ontario Renters	0.0804	7.35
	Quebec Renters	0.0255	2.66
	New Brunswick Renters	0.1064	5.61
	Yukon Renters	0.0639	3.80
	Northwest Territories Renters	0.0682	6.22

The final regression model then had various diagnostic tests performed on it. Since the regression is a weighted least squares with a random error term, Lange and Ryan (1989) have suggested using the following form to create standardized residuals:

$$z_i = \frac{\hat{F}_i^{(eb)} - X_i \hat{\beta}}{\sqrt{\sigma_i^2 + \hat{\tau}^2}}.$$

The residuals were analyzed using both Q-Q plots and outlier detections methods: no major departures from the assumed distribution of the residuals were detected. More details on the residual analysis can be found in Dick (1993).

Table 2
Direct, Smoothed and Raked Estimates of Adjustment Factors

Sex	Age	Estimate	B.C.	Alta	Sask.	Man.	Ont.	Que.	N.B.	N.S.	P.E.I.	Nfld	Yukon	N.W.T.
Male	0-19	Direct	1.017	1.026	1.012	1.029	1.028	1.017	1.022	1.019	1.004	0.999	1.031	1.036
		Smooth	1.019	1.013	1.009	1.013	1.029	1.016	1.027	1.010	1.007	1.006	1.026	1.027
		Raked	1.020	1.016	1.011	1.015	1.031	1.018	1.027	1.013	1.005	1.007	1.029	1.031
	20-29	Direct	1.087	1.036	1.068	1.058	1.113	1.071	1.122	1.063	1.060	1.057	1.098	1.127
		Smooth	1.086	1.056	1.065	1.062	1.104	1.074	1.103	1.064	1.063	1.062	1.094	1.122
		Raked	1.083	1.061	1.073	1.067	1.101	1.079	1.096	1.073	1.041	1.074	1.096	1.127
	30-44	Direct	1.031	1.021	1.028	1.034	1.054	1.047	1.043	1.018	1.025	1.026	1.069	1.080
		Smooth	1.039	1.026	1.028	1.030	1.053	1.041	1.046	1.026	1.028	1.028	1.052	1.059
		Raked	1.038	1.028	1.032	1.032	1.051	1.043	1.043	1.029	1.018	1.033	1.053	1.059
	45 +	Direct	1.019	1.018	1.002	1.014	1.013	1.011	1.014	1.016	1.018	1.016	0.992	1.076
		Smooth	1.017	1.011	1.006	1.009	1.019	1.013	1.019	1.010	1.009	1.009	1.021	1.039
		Raked	1.014	1.010	1.006	1.009	1.016	1.012	1.015	1.010	1.005	1.010	1.019	1.035
Female	0-19	Direct	1.034	1.018	1.017	1.012	1.037	1.029	1.029	1.014	0.995	1.016	1.026	1.054
		Smooth	1.030	1.015	1.013	1.015	1.038	1.023	1.030	1.010	1.006	1.010	1.028	1.061
		Raked	1.032	1.018	1.016	1.017	1.040	1.026	1.030	1.012	1.004	1.013	1.030	1.068
	20-29	Direct	1.068	1.047	1.028	1.020	1.072	1.043	1.070	1.030	1.004	1.041	1.068	1.072
		Smooth	1.058	1.036	1.031	1.029	1.070	1.044	1.071	1.031	1.027	1.033	1.069	1.092
		Raked	1.058	1.041	1.036	1.032	1.070	1.048	1.068	1.037	1.018	1.041	1.072	1.099
	30-44	Direct	1.013	1.009	1.004	1.006	1.027	1.017	1.031	1.019	1.004	1.024	1.031	1.020
		Smooth	1.018	1.008	1.007	1.007	1.030	1.017	1.029	1.010	1.007	1.011	1.028	1.026
		Raked	1.017	1.008	1.007	1.007	1.028	1.017	1.025	1.011	1.004	1.012	1.027	1.026
	45 +	Direct	1.007	1.003	1.018	1.001	1.011	1.011	1.000	1.002	0.993	1.013	1.024	1.007
		Smooth	1.014	1.006	1.010	1.006	1.021	1.015	1.020	1.006	1.005	1.009	1.031	1.026
		Raked	1.008	1.004	1.007	1.004	1.012	1.009	1.011	1.004	1.002	1.006	1.019	1.016

3.3 Estimates of Adjustment Factors

Table 2 shows both the direct survey estimate and the smoothed Empirical Bayes estimate of the adjustment factors. An inspection of the table shows that these estimates are relatively close, reflecting the Empirical Bayes methodology of combining the direct survey estimate with the model estimate. Note that all of the domains that were originally estimated to have overcoverage – shown by an estimated adjustment factors being less than one – have been changed, by the Empirical Bayes estimates to being an estimate of undercoverage. The difference between the two sets of estimated adjustment factors – in absolute terms – differ by under 1% and in the larger provinces by less than 0.5%. However, for some of the smaller provinces and territories the difference between the two estimates can be substantially larger. In the Northwest Territories the change between the directly estimated adjustment factor and the Empirical Bayes estimate is about 2% for 3 age – sex groups and over 3% for another.

The objective of the Empirical Bayes model is to produce estimates with smaller MSE than the survey estimates. From Section 2.2 it can be shown that the variance for the direct survey estimates is calculated from

log v̂(F_i) = - 6.133 - 0.285 log C_i,

while the Prasad-Rao MSE, from Section 2.3, is calculated by

MSE[F̂_i^{eb}] = MSE(F̃_i^{eb}) + 2ĝ_{3i}(τ̂²) =
ĝ_{1i}(τ̂²) + ĝ_{2i}(τ̂²) + 2ĝ_{3i}(τ̂²).

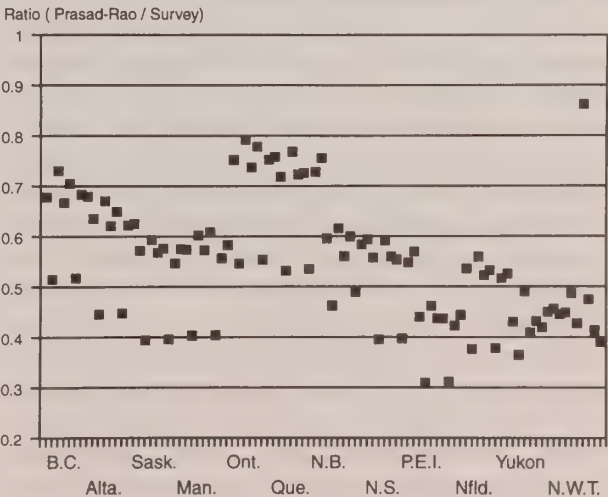


Figure 2. Ratio of root MSE, Prasad-Rao and survey.

Figure 2 plots, for each domain, $R = \sqrt{\widehat{\text{MSE}}[\hat{F}_i^{\text{eb}}] / \hat{v}(F_i)}$ the ratio of the root mean square errors for the Empirical Bayes model and the estimated survey variance (Note that within provinces the domains are ordered as Male aged 0-19, 20-29, 30-44 and 45 and over and Female aged 0-19, 20-29, 30-44 and 45 and over). Clearly, the Empirical Bayes MSE is smaller in all domains. However, in the larger provinces, Ontario and Quebec, the ratio of the root MSEs is only between 0.7 and 0.8. This relatively small gain is a reflection of the large sample sizes in these domains which in turn give a reliable estimate of the variance. The large gains are made in the smaller provinces and territories. For instance, in Prince Edward Island, the ratio of the root MSEs are all smaller than 0.5 showing the large improvement in the estimates. The one outlier is in the Northwest Territories (females aged 0 to 19): the Prasad-Rao MSE appears to have been overestimated in this domain.

4. ADJUSTMENTS MADE TO EMPIRICAL BAYES ESTIMATES

4.1 Rationale and Methodology

The advantage of the Empirical Bayes method is apparent from the above discussion. However, the Empirical Bayes methodology does not preserve the higher level (*i.e.*, the large domain) direct survey estimates that are reliable. By this it is meant that the provincial totals and the age – sex domain totals for the direct survey estimates and the Empirical Bayes estimates are not equal. Since the two surveys were designed to produce estimates at these levels, it is crucial that the Empirical Bayes be consistent with these reliable marginal totals.

To achieve consistency of estimates of missed persons between the reliable provincial and age – sex totals from the direct survey estimates and the final Empirical Bayes estimates, a raking ratio procedure was used. This is basically the method used in Australia to determine their small domain estimates (see Choi *et al.* 1988). This technique re-scales the individual Empirical Bayes estimates to conform to the known provincial and national age – sex totals. Once this procedure has converged, the final estimates will be consistent with the direct survey totals. In terms of a log-linear model, we are using as the main effects (province and age-sex) estimates the results from the two coverage studies and the interaction terms (province by age-sex) estimates from the Empirical Bayes modelling.

Details of the procedure can be described as follows. Assume that we have a matrix of estimated missed persons that has P columns (corresponding to the provinces) and A rows (corresponding to the age-sex groups). First set $F_{pa} = F_i$, then let $\hat{M}_{pa}^s = C_{pa}(F_{pa} - 1)$ be the direct survey estimate of the number of missed persons in province p and age – sex group a and let $\hat{M}_{pa}^{(\text{eb})} = \hat{M}_{pa}^{(0)} = C_{pa}(\hat{F}_{pa}^{(\text{eb})} - 1)$ be the Empirical Bayes estimate of missed persons from

the Empirical Bayes model. If we let a plus sign (+) represent addition across the variable then the raking estimate can be written for cycles $\kappa = 0, 1, \dots$ as;

$$\hat{M}_{pa}^{(2\kappa+1)} = \hat{M}_{pa}^{(2\kappa)} \left(\sum_{a=1}^A \hat{M}_{pa}^s / \sum_{a=1}^A \hat{M}_{pa}^{(2\kappa)} \right)$$

and

$$\hat{M}_{pa}^{(2\kappa+2)} = \hat{M}_{pa}^{(2\kappa+1)} \left(\sum_{p=1}^P \hat{M}_{pa}^s / \sum_{p=1}^P \hat{M}_{pa}^{(2\kappa+1)} \right).$$

This procedure will converge to a unique solution. Since this is basically a log-linear model, the underlying assumption is that the relationship determined by the Empirical Bayes model for the interaction between province and age – sex group is valid and will be preserved.

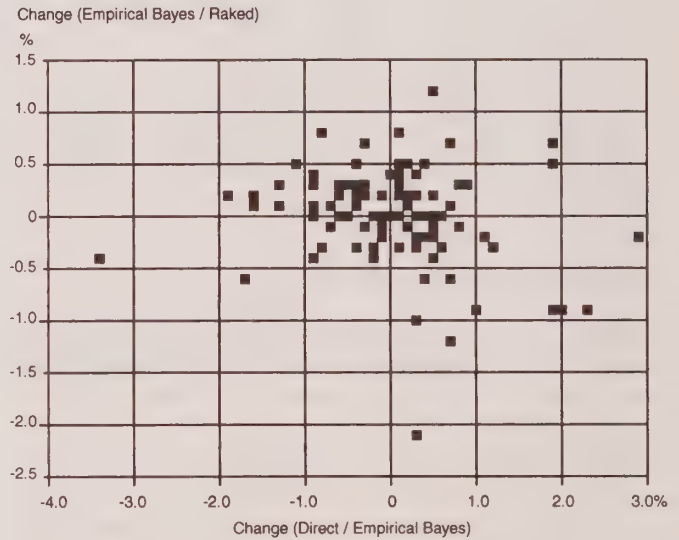


Figure 3. Percent change in estimates of adjustment factors.

Table 2 shows the final raked estimates of the adjustment factors along with both the original survey estimates and the Empirical Bayes estimates. Generally, the impact of raking is to shrink the Empirical Bayes estimate back towards the survey estimate. This is shown in Figure 3. Here two different percent changes in the estimated adjustment factors are plotted. The X-axis shows the percent change between the direct survey estimate and the Empirical Bayes estimate. The Y-axis shows the percent change between the Empirical Bayes estimate and final raked estimate. The plot shows that the two variables are negatively correlated: hence the raking tends to move the Empirical Bayes estimates closer to the original survey estimates.

One draw back of this procedure is that the MSEs of the raked adjustment factors are now very difficult to estimate. Due to the non-linear nature of the raking ratio procedure, a direct calculation is impossible. It is possible to use a Taylor series expansion; however this assumes a large sample size in each domain when in fact we know some domains have very small sample sizes. A possible procedure is to adjust the estimated MSE from the Empirical Bayes estimates and multiply these by the squared ratio of the raked Empirical Bayes estimate over the Empirical Bayes estimate. While this procedure is only a crude approximation, it can at least give some guidance as to the reliability of the individual estimates. This method will ensure that the coefficient of variations calculated for the Empirical Bayes estimates will be retained for the corresponding raked Empirical Bayes estimates. This is the procedure that was used to produce the final MSE estimates for the raked Empirical Bayes estimates of missed persons.

4.2 Detailed Domain Estimates

The Population Estimates Program requires even finer detail than that produced by the various models discussed above. In fact the program needs estimates for single years of age for each sex for each Census Division within each province. Since the Empirical Bayes methodology is limited somewhat by the direct survey results – an estimate with a non-zero standard error is required for each domain – synthetic methods must be used to generate the more detailed estimates.

For the Population Estimates Program, estimates for each province and sex were produced for 9 age groups instead of the 4 age groups used in the Empirical Bayes model. A straight synthetic model, using the raked Empirical Bayes estimates as initial values, was proposed for this stage of estimation. To produce these more detailed estimates, the raked Empirical Bayes estimate was allocated proportionally by Census count across all sub-age groups within each province and sex. Let the final raked estimate in the p -province and the a -th age-sex group be $\hat{M}_{pa}^{2k+2} = \hat{M}_{pa}^{rf}$. Also if the a -th age – sex group is composed of Q exclusive sub-age groups then the estimate of the missed persons in the p -th province and the q -th sub-age group within the a -th age – sex group would be

$$\hat{M}_{paq} = \hat{M}_{pa}^{rf} \left(\frac{C_{paq}}{C_{pa+}} \right),$$

where $C_{pa+} = C_{pa} = \sum_{q=1}^Q C_{paq}$. This approach guarantees that the estimates from the earlier raked Empirical Bayes output are preserved for the original domain total. The further estimates that are required for the population estimates program use demographic methods. In fact, one

of the objectives of the Empirical Bayes procedure is to provide initial estimates for the demographic methods. See Michalowski (1993) for further details.

5. SUMMARY AND CONCLUSIONS

The Empirical Bayes methodology was adopted because it preserves the more reliable estimates from the larger provinces and domains while permitting a model based estimate to dominate if the underlying direct estimate is unreliable. This is in accordance with standard survey methods of using the direct survey estimates as much as possible. The raking ratio procedure used for adjusting the estimates from the Empirical Bayes model was used to ensure consistency with the direct survey results that were known to be reliable.

As for the explicit model used to describe the underlying true adjustment factors, it must be noted that this model is purely descriptive. Its primary function is to use explanatory variables to describe the variation in adjustment factors, taking into account the sampling error associated with each adjustment factor. It would not be prudent to make far-reaching conclusions on the nature of under-coverage from the final set of parameters included in the model.

The main weakness of this approach is with the two variances that are estimated. The assumption of the regression model errors being approximately normally distributed is difficult to assess. In the absence of any real knowledge about the true underlying distributions any assumption about the model variance will be essentially unverifiable. The proposed model variance seems reasonable and diagnostic checks have not revealed any major problems.

The sampling variance model is more problematic. All Empirical Bayes methods assume that this variance is known, when in fact it has to be estimated. Efforts to extend the Prasad-Rao MSE calculation to include the contribution from this estimated parameter have not yielded any new results.

In the future, research will concentrate in working around the problem associated with estimating the sampling variances. Further work needs to be conducted on the Prasad-Rao MSE calculation. In addition, the possibility of using the micro level data from the coverage studies and estimating the undercoverage rates directly through logistic regressions as in Wong and Mason (1985) will be pursued.

Another project would be to examine the implications of recasting the Empirical Bayes model into the standard state space framework (Robinson 1991). Pfeiffermann and Burck (1990) have suggested a method for calculating the MSE for a time series placed in a state space model that has to conform to certain periodic benchmarks. The state space formulation would also be useful in explicitly incorporating the demographic methods.

ACKNOWLEDGEMENTS

The author is grateful to D. Binder, M. Hidirolou, R. Carter, M. Armstrong, J. Tourigny, J.N.K. Rao and especially D. Royce for commenting on an earlier version of this paper. An Associate Editor and two referees also provided comments that improved the final version.

REFERENCES

- BURGESS, R.D. (1988). Evaluation of Reverse Record Check estimates of undercoverage in the Canadian Census of Population. *Survey Methodology*, 14, 137-156.
- CHOI, C.Y., STEEL, D.G., and SKINNER, T.J. (1988). Adjusting the 1986 Australian Census count for underenumeration. *Survey Methodology*, 14 173-189.
- CRESSIE, N. (1992). REML estimation in Empirical Bayes smoothing of census undercount. *Survey Methodology*, 18, 75-94.
- DATTA, G.S., GHOSH, M., HUANG, E.T., ISAKI, C.T., SCHULTZ, L.K., and TSAY, J.H. (1992). Hierarchical and Empirical Bayes methods for adjustment of census undercount: The 1988 Missouri Dress Rehearsal data. *Survey Methodology*, 18, 95-108.
- DICK, J.P. (1993). Procedures used in modelling net undercoverage in the 1991 Census. Internal Statistics Canada memorandum.
- DRAPER, N.R., and SMITH, H. (1966). *Applied Regression Analysis*. New York: John Wiley and Sons.
- ERICKSEN, E.P., and KADANE, J.B. (1985). Estimating the population in a census year (with discussion). *Journal of the American Statistical Association*, 84, 927-943.
- FAY, R.E., and HERRIOT, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 82, 269-277.
- FREEDMAN, D., and NAVIDI, W. (1992). Should we have adjusted the U.S. Census of 1980? (with discussion). *Survey Methodology*, 18, 3-74.
- GERMAIN, M.-F., and JULIEN, C. (1993). Results of the 1991 Census coverage error measurement program. *Proceedings of Seventh Annual Research Conference*. United States Bureau of the Census, 55-70.
- GHOSH, M., and RAO, J.N.K. (1994). Small area estimation: An appraisal (with discussion). *Statistical Science*, 9, 55-93.
- HARVILLE, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72, 320-337.
- HOGAN, H. (1992). The 1990 Post-Enumeration Survey: an overview. *The American Statistician*, 46, 261-269.
- JUDGE, G.G., GRIFFITHS, W.E., CARTER HILL, R., and LEE, T-C. (1984). *The Theory and Practice of Econometrics*. New York: John Wiley and Sons.
- LANGE, N., and RYAN, L. (1989). Assessing normality in random effects models. *The Annals of Statistics*, 17, 624-642.
- MARITZ, J.S., and LWIN, T. (1989). *Empirical Bayes Methods (2nd edition)*. London: Chapman and Hall.
- MICHALOWSKI, M. (1993). Revised postcensal and intercensal estimates: Canada, provinces and territories, 1971 - 1991. Internal report, Population Estimates Section, Statistics Canada.
- PFEFFERMANN, D., and BURCK, L. (1990). Robust small area estimation combining time series and cross-sectional data. *Survey Methodology*, 16, 217-237.
- PRASAD, N.G.N., and RAO, J.N.K. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, 85, 163-171.
- RAO, C.R. (1973). *Linear Statistical Inference and its Applications*. New York: John Wiley and Sons.
- ROBINSON, G.K. (1991). That BLUP is a good thing: the estimation of random effects (with discussion). *Statistical Sciences*, 6, 15-51.
- ROYCE, D. (1992). A comparison of some estimators of a set of population totals. *Survey Methodology*, 18, 109-125.
- STATISTICS CANADA (1993). *1991 Census Technical Report: Coverage*. Ottawa: Supply and Services Canada, 1994. 1991 Census of Canada: Catalogue No. 92-341E.
- WONG, G.Y., and MASON, W.M. (1985). The hierarchical logistic regression model for multilevel analysis. *Journal of the American Statistical Association*, 80, 513-524.

Between-State Heterogeneity of Undercount Rates and Surrogate Variables in the 1990 U.S. Census

JAY JONG-IK KIM, ALAN ZASLAVSKY and ROBERT BLODGETT¹

ABSTRACT

As part of the decision on adjustment of the 1990 Decennial Census, the U.S. Census Bureau investigated possible heterogeneity of undercount rates between parts of different states falling in the same adjustment cell or poststratum. Five “surrogate variables” believed to be associated with undercount were analyzed using a large extract from the census and significant heterogeneity was found. Analysis of Post Enumeration Survey on undercount rates showed that more variance was explained by poststratification variables than by state, supporting the decision to use the poststratum as the adjustment cell. Significant interstate heterogeneity was found in 19 out of 99 poststratum groups (mainly in nonurban areas), but there was little if any evidence that the poststratified estimator was biased against particular states after aggregating across poststrata. Nonetheless, this issue should be addressed in future coverage evaluation studies.

KEY WORDS: Poststratification; Influence statistics; Linearization; Synthetic estimation.

1. INTRODUCTION

The Post Enumeration Survey (PES) of the 1990 Decennial Census of the United States was designed to produce coverage estimates for 1,392 poststrata. The nation was first divided into 116 domains, called poststratum groups (PSGs) according to geography, race/Spanish origin and tenure (owner vs. renter). With only 4 exceptions, all PSGs are defined within a census division, one of nine contiguous geographic areas each composed of several states. Each PSG was further divided into 12 age-by-sex groups, the poststrata. For example, roughly all Black renters in New York city constitute a PSG and all females, age 0-9, of this PSG make a poststratum (PS). Further details on the PES are in Hogan (1992,1993).

Small area undercount rates were calculated by synthetic estimation; the same adjustment factor was applied to persons from a given PS in all areas. This procedure is accurate under the “synthetic assumption” of homogeneity of undercount rate within a PS. The validity of the synthetic assumption has been hotly debated (Section 2). This paper reports on research conducted as a part of a PES evaluation project (the “P12 project”) which investigated heterogeneity within poststrata. In particular, this research focused on the following question: can differences in coverage be identified between parts of a poststratum that fall into different states?

Under the homogeneity assumption, the rates are the same within a PS regardless of state. Thus, this assumption can be tested by comparing rates from state to state within a PS; this test focuses attention on the question of whether synthetic estimation is “unfair” to certain states. The unit

of analysis is the intersection of a census block and a PS or PSG, called a block part (BP) for the analysis of the undercount rate data. A census block is a small area bounded by visible features such as streets, streams *etc.* and/or by political boundaries. In urban areas it roughly corresponds to a city block. In fact, most of our analyses are performed on PSGs, since the age-sex breakdown of the PSG did not vary much from state to state. Hence, the analysis focuses on whether BPs differ between states within PSG.

Two distinct analyses were performed. The distributions of five “surrogate variables” were investigated (Section 3), using a large (4.26%) extract from the census. The distribution of undercount was investigated using the much smaller PES data set (Section 4). For more detailed tables and documentation of the project, see Kim (1991).

2. LITERATURE REVIEW

Two key questions have been addressed in the literature on heterogeneity:

1. The empirical question: how much heterogeneity is there, and how can it be described?
2. The theoretical and policy question: what are the implications of heterogeneity for the accuracy of synthetic adjustments and the validity of assessments of these adjustments?

Heterogeneity may be identified and analyzed at many levels of aggregation. Perfect homogeneity of undercount rates for very small domains is numerically impossible,

¹ Jay Jong-Ik Kim, Statistical Research Division, U.S. Bureau of the Census, Suitland, MD 20233, U.S.A.; Alan Zaslavsky, Department of Statistics, Harvard University, Cambridge, MD 02138, U.S.A.; and Robert Blodgett, U.S. Food and Drug Administration, 200 C St., S.W., Washington, DC 20204, U.S.A.

because of discreteness of the true population and the census counts. Indeed, because census errors (omissions or erroneous enumerations) tend to be either independent of each other or positively associated (as when a household with several members is omitted, or when some local characteristic affects an entire block), we would anticipate at least binomial variability in observed undercount rates.

Hengartner and Speed (1993) analyzed 1990 PES data from two sites by fitting models in which the explanatory variables were block and “demoid” (a unit defined by the non-geographic poststratification variables, such as race, sex, age, and tenure). They found that the amount of variance explained by block was slightly greater than the amount explained by demoid; the number of blocks was not much greater than the number of demoids in their data set. In response, Schafer (1993) argued that an estimation scheme involving block effects would not be practical because it would require collecting data from every block.

Heterogeneity of undercount at any level may be defined as excess variability in observed undercount rates at that level over what would be expected as a consequence of variability at a lower level of aggregation. For example, confining our attention to a single poststratum, a set of blocks are heterogeneous if their undercount rates in that poststratum differ more than would be expected if households, including those counted, partially counted, and omitted in the census, had been randomly distributed across the blocks. Similarly, a group of states are heterogeneous (similarly controlling for poststratum) if they differ more than would be expected if blocks, including those with higher and lower undercounts, had been randomly distributed across the states. Several studies have attempted to measure heterogeneity in undercount rates and other census variables. Wachter and Freedman (1992) analyzed a large sample of census data (similar to that considered in Section 3). They estimated the excess variability between “superblocks” over that predicted by a binomial model with uniform rates, for four “artificial population” variables (multi-unit housing rate, non-mailback rate, allocations, and substitutions, described in Section 3). Compared to the greatest possible amount of heterogeneity (if each block were homogeneous), the “excess variability” ranged from around 20% (for multi-unit housing) to 2% (for substitutions). Another study by Freedman and Wachter (1993) examined between-state heterogeneity using “artificial populations” based on the same variables and two others, and found substantial variability.

Alho, Mulry, Wurdeman and Kim (1993) used conditional logistic regression models to describe heterogeneity associated with measured covariates that were not captured in the poststratification. Their concern was primarily with reducing the bias of dual system estimates of population, rather than with more accurate small-area estimates.

A controversial topic in evaluation of the proposed adjustment of the 1990 census was the effect of heterogeneity on the accuracy of adjusted population counts obtained by synthetic estimation, and particularly on comparisons of the accuracy of adjusted and unadjusted counts. Wachter and Freedman (1992) argued that because the “synthetic assumption” of uniform coverage within poststrata is demonstrably false, aggregate measures of the accuracy of an adjusted census systematically underestimate error. Because nonuniformity of coverage affects the accuracy of an unadjusted census as well, however, the implications of this conclusion for the appropriateness of adjustment are not obvious.

In one of the earlier “surrogate variables” studies, Isaki, Schultz, Diffendal and Huang (1988) simulated the behavior of synthetic estimators on “artificial populations” which were transformations of the substitution (unit imputation) rate. They found that a synthetic estimator generally did better than “unadjusted” counts.

Schirm and Preston (1987) argued, using analytical calculations and simulation, that synthetic estimation makes estimates for small areas more accurate under plausible conditions, even if the synthetic assumption does not hold. Wolter and Causey (1991) investigated the performance of synthetic estimators and of a single ratio adjustment when the undercount rates are estimated with error, using undercount rates from the 1980 Post-Enumeration Program (PEP) and simulating various levels of sampling error; they estimated “break-even” coefficients of variation at which sampling error in the adjusted counts or proportions would make them less accurate than unadjusted counts or proportions. The conclusions of these articles were criticized by Freedman and Navidi (1992), who gave counterexamples of possible distributions of undercount for which adjustment by synthetic estimation would make population distribution less accurate.

Fay and Thompson (1993) simulated effects of heterogeneity on accuracy of synthetic estimates, using eight surrogate variables (including the five used in this study) and the same data set as analyzed in Section 3. They performed a loss function analysis as in Mulry and Spencer (1993) to compare the accuracy of simulated unadjusted counts to that of synthetically adjusted counts. They found that the effect of ignoring heterogeneity was to underestimate the gain in accuracy due to synthetic adjustment for five of eight variables, and to overestimate it for one variable (unemployment rate), while there was little difference for two other variables (poverty and migration rates).

3. ANALYSIS OF SURROGATE VARIABLES

In the analysis of census data, we selected variables which were available for the entire census and which, like undercount, were descriptive of or related to the

census-taking process. The selected surrogates are the allocation rate, mail return rate, multiunit structure rate, mail universe rate (fraction of units receiving mail questionnaire) and substitution rate. The allocation rate is the fraction of households for which imputations were made for item nonresponse, and the substitution rate is the fraction of households which were imputed as a whole because it was determined that a unit was occupied but no interview could be obtained.

Table 1 shows correlations between each of these variables and undercount rate by PSG. These "ecological" correlations (Freedman, Pisani and Purvis 1978, pp. 141-142) of PSG averages differ from those which could be calculated from block-level data. The latter are smaller, possibly because of the noise introduced by random variability in the small populations in each block.

Table 1

Correlation Coefficients between the
Surrogate Variable
and Undercount Rate by PSG

Variable	Correlation
Allocation Rate	.44
Mail Return Rate	-.57
Multiunit Structure Rate	.39
Mail Universe Rate	.08
Substitution Rate	.47

Applying a naive test which treats the PSGs as independent, each correlation is significant except that for mail universe rate, but the magnitudes of the correlations are not large. To some extent, furthermore, these variables are descriptive of conditions which tend to lead to higher omission rates (allocations due to poor completion of questionnaires, substitutions due to difficulty in obtaining interviews) or to lower omission rates (high mail return rates). On the other hand, difficult census-taking conditions can also lead to erroneous enumerations, so these effects on net undercount are not entirely clear-cut. We do not analyze these variables simply because we believe that they are distributed in exactly the same way as undercount. Rather we hope that by obtaining results on the distributions of a range of different census variables, we may gain some insight into the distribution of undercount.

For the analyses of the surrogate variables, a stratified cluster sample of 1990 Census data was extracted. This sample is composed of 204,394 blocks corresponding to 125,000 block clusters. A block part containing less than ten persons was combined with successive block parts (in order by block number) until a minimum count of ten persons was obtained. This operation was performed to obtain relatively stable rates for the surrogate variables which allows us to analyze the rates themselves.

Surrogate variables are analyzed by logistic regression. Two forms of logistic regression model were used. For the within-PSG analysis, the model for PSG i is

$$\log [P_{ij} / (1 - P_{ij})] = A + C_j$$

and for the within-division analysis,

$$\log [P_{ij} / (1 - P_{ij})] = A + B_i + C_j,$$

where P_{ij} is the rate for a surrogate variable in the i -th PSG and j -th state, A is the intercept, B_i is the i -th PSG effect and C_j is the j -th state effect. The models used only the 99 PSGs astride two or more states. Models were built for surrogate variables in the 99 PSGs and in each of nine divisions. SAS PROC CATMOD estimated the parameters by maximum likelihood and provided Wald statistics for testing the significance of state effects.

The data were collected with a cluster sample rather than a simple random sample so the test statistics must be divided by a design effect. We estimate a design effect,

$$\hat{D}_{ij} = \frac{\sum_{k=1}^{K_{ij}} n_{ijk} (\hat{p}_{ijk} - \hat{p}_{ij})^2}{K_{ij} \hat{p}_{ij} (1 - \hat{p}_{ij})},$$

where \hat{p}_{ijk} is the rate for the i -th PSG, j -th state and k -th combined BP; n_{ijk} is the size of the combined BP; K_{ij} is the sample number of combined BPs in the i -th PSG in the j -th state and \hat{p}_{ij} is the estimated rate for the i -th PSG and j -th state. The fraction is the ratio of the observed between-block variance to that expected under binomial sampling.

The estimated design effect \hat{D}_{ij} is a measure of within-state within-PSG heterogeneity. The more within-state heterogeneity there is, the greater the sampling variance of the state-level rate and the harder it is to detect a significant difference. The magnitude of the design effect thus affects the statistical power of the hypothesis tests.

The calculated design effect only approximates the required correction. First, \hat{D}_{ij} sums over the combined BPs rather than individual BPs. Second, the sample is a stratified cluster sample, and most or all post-strata span several sampling strata. The formula is only strictly correct for an unstratified sample. Third, the correct effect involves off-diagonal (covariance) as well as on-diagonal (variance) terms, and the off-diagonal terms are omitted. To account for the above, the calculated design effects were multiplied by the judgmentally chosen factor, 1.25.

A design effect was calculated for each surrogate variable and PSG. It is small (around 2) in most PSGs for the allocation and substitution rate. The effect is slightly higher for mail return rate, but it tends to be large (as much as 20) for multiunit structure and mail universe rate, since these characteristics are usually fairly uniform within a block but vary greatly between blocks.

Table 2 summarizes the design-corrected tests for state effects within PSG.

Table 2
Number of PSGs with Significant ($\alpha = .05$)
State Effect (Logistic Regression)

Div.	No. Grp	Alloc	Mail Ret	Mult Str	Mail Unv	Sub
1	5	5	5	5	1(1)	3(4)
2	12	11	11	12	7(10)	12
3	16	15	16	16	3(3)	12(12)
4	8	8	8	7	5(6)	5(8)
5	10	10	9	10	4(4)	7(8)
6	15	15	13	15	5(7)	15
7	9	8	9	9	4(4)	8(8)
8	7	7	7	7	2(3)	6(6)
9	17	15	14	14	5(5)	6(12)
Sum	99	94	92	95	36(43)	74(84)

The numbers in () are the number of PSGs for which test statistics are available when they are less than the number of groups.

Nationally, for each surrogate variable the state effect is significant for at least 84% of the PSGs. (The total number of PSGs varies because when a PSG falls entirely within one state or when only one state has non-zero observations for a particular variable, the corresponding model cannot be fit). The results are summarized at the division level. (Divisions 1 through 9 are New England, Mid-Atlantic, South Atlantic, East South Central, West South Central, East North Central, West North Central, Mountain and Pacific Divisions.)

Table 3 shows the magnitude of state effects, expressed as χ^2 values of test statistics adjusted for design effect, for three variables having relatively high correlation with the undercount rate. In the table, the χ^2 values have from 1 to 8 degrees of freedom.

Table 3
Magnitude of State Effects with respect to
Test Statistics

	Allocation Rate	Mail Return Rate	Substitution Rate
Minimum	4.3	0.28	5.46
25%-ile	27.5	102.83	49.80
50%-ile	68.9	254.49	97.35
75%-ile	140.3	644.05	260.88
Maximum	945.2	8,779.88	1,815.12

In division-level models with state and PSG effects, both the state and PSG effects were significant at the 1% level in every division and for every variable (excluding mail universe rate in two divisions where a test statistic could not be calculated).

4. ANALYSIS OF UNDERCOUNT RATE

The results described above for surrogate variables were obtained early in the census process, but they have limited relevance to homogeneity of undercount itself. After PES data were processed, direct analysis of the distribution of undercount became possible.

The data set for these analyses merged two data sets for the 12,124 PES sample blocks, one for the *E*-sample (Census follow-up) and the other for the *P*-sample (PES). There were 12,124 collection blocks, some of which were split up for tabulation, giving 12,964 tabulation blocks. More importantly, because some of the smaller blocks were combined in the sampling, there were 5,293 block clusters sampled. Correct enumerations and *E*-sample total counts are on the *E*-sample file. The *P*-sample file includes *P*-sample total counts and counts of matches (*P*-sample cases that were included in the Census).

4.1 Variance Explained by State and PSG

For each division, a two-way ANOVA was fitted to undercount rates for state parts. Table 4 shows the ratio of the sum of squares due to PSGs to that due to states within a division.

Table 4
Variance of Undercount Rate Explained
by State and PSG

Div.	No. of Groups	No. of States*	SS (Group) SS (State)	MS (Group) MS (State)
1	5	6	4.51	5.64
2	12	3	4.88	.89
3	16	9	12.69	6.77
4	8	4	8.73	3.74
5	10	4	8.17	2.72
6	15	5	7.67	2.19
7	9	7	2.78	2.09
8	7	8	1.31	1.53
9	17	5	40.28	10.07

* States include D.C.

The ratio is always greater than one and in Division 9 it is 40.28, showing much larger effects for PSG than for state. The mean square for group also exceeds the mean square for state in each division except Division 2. This supports the decision to use the PS rather than the state as the cell for undercount estimation and adjustment.

4.2 Tests for State Effects on Undercount Rates

Assuming the substitution rate (fraction of units imputed for nonresponse) is negligible, the adjustment factor (\hat{R}) for a domain is

$$\hat{R} = \frac{WCE/WE}{WM/WP},$$

and the undercount rate is

$$1 - 1/\hat{R},$$

where WE and WP are the estimated population sizes weighted up from the E and P -sample, respectively. WCE is the weighted number of correct enumerations and WM is the weighted number of matches in the PES.

The statistic for the influence (see Appendix) of the i -th BP on the adjustment factor or undercount rate is

$$I_i = \hat{R} \left(\frac{WCE_i}{WCE} + \frac{WP_i}{WP} - \frac{WE_i}{WE} - \frac{WM_i}{WM} \right),$$

where WCE_i , WP_i , WE_i and WM_i are contributions from the i -th BP to the totals above.

A linear model was fitted to BP influence statistics to test for state effects. Under the null hypothesis, all the state parts in a PSG have the same undercount rate and the expected mean of the influence statistics for each state is 0 within each PSG. The influence statistics can be analyzed with one way ANOVA within a single PSG or two way ANOVA for all PSGs within a division.

Table 5 summarizes the tests for state effects on linearized statistics within each PSG.

Table 5

Analysis of Linearized Undercount at the PSG Level

Division	Number of PSG	Number of PSG with $P < .05$
1	5	0
2	12	3
3	16	4
4	8	5
5	10	2
6	15	1
7	9	0
8	7	1
9	17	3
Sum	99	19

The tests reveal significant heterogeneity between states in 19 out of 99 groups at the 5% significance level. The magnitude of the estimated state effect ranges from a few percent up to 20%, but the standard errors of these estimates are very large.

Table 6 summarizes the results of these analysis by place type. Place types 0, 1, 2 and 3 are large central cities in a Primary Metropolitan Statistical Area (PMSA), place types 4, 5 and 6 are non-central cities in PMSA with large central cities and place types 7, 8 and 9 are other areas.

The significant results are concentrated in PSGs for small areas (place types 7, 8 and 9). Ten out of 32 such

groups show significant interstate heterogeneity at the 5% level. This suggests that the poststratification can be improved in those areas.

Table 6

Summary of Analysis of Linearized Undercount by Place Type

Place Type	Number of PSG	Number of PSG with $P < .05$
0	11	3
1	23	1
2	12	1
3	8	1
4	0	0
5	6	2
6	6	1
7	11	3
8	11	4
9	10	3

Table 7 shows the F -statistics and p -value for state effect for state \times PSG models, once weighted by the size of domain and once without weights.

Table 7

State Effects by Division – Weighted and Unweighted Data

Division	D.F.	Unweighted Models		Weighted Models	
		F	p	F	p
1	5	.57	.72	.40	.85
2	2	4.64	.01	1.72	.18
3	8	.43	.91	.65	.74
4	3	.64	.59	.66	.58
5	3	.66	.58	1.37	.25
6	4	.60	.66	.24	.92
7	6	.39	.88	.22	.97
8	7	.62	.74	.76	.62
9	4	.77	.54	.48	.75

The additive effect of state was significant in only one division ($p = .01$) in the unweighted state \times PSG model; when data were weighted by size of domain, the smallest p -value for the state effect was .18. In both cases, the most significant effect was observed in Division 2, in which New Jersey appeared to have higher undercount rate, controlling for PSG, than New York. Note that the most undercounted area in New York (New York City) had its own poststrata. In eight out of ten PSGs for which New Jersey and New York could be compared, including nonurban areas, the estimated undercount for New Jersey was larger than that for New York. Elsewhere, because the state effects in

different PSGs varied in magnitude and sometimes in sign, and because only within a minority of PSGs in any division were there significant state effects, there was not significant evidence that in the aggregate the poststratification was biased against certain states.

Table 8 shows point estimates of the state effects in linear models for undercount rate by state part in each division, with effects for state and poststratum group. (Effects are centered at zero by division.) In effect, these are estimates of interstate differences after correcting for effects explained by the PSG composition of the different states.

Table 8

Estimated State Effects on Undercount within Division
(as percent)

Division 1		Division 4		Division 7	
CT	-2.42	AL	-2.90	IA	-1.10
ME	.74	KY	1.89	KS	-0.50
MA	-0.48	MS	-0.02	MN	-0.01
NH	-0.14	TN	1.03	MO	-0.66
RI	1.43			NE	1.76
VT	0.90			ND	-0.07
				SD	0.60
Division 2		Division 5		Division 8	
NJ	4.18	AR	1.44	AZ	2.70
NY	-3.91	LA	-0.71	CO	0.68
PA	-0.26	OK	1.58	ID	-2.24
		TX	-2.30	MT	-1.61
Division 3				NV	-0.10
DE	-0.42			NM	3.35
DC	2.82			UT	0.08
FL	-0.88			WY	-2.84
GA	-1.43	Division 6		Division 9	
MD	-1.32	IL	0.86	AK	-0.78
NC	0.53	IN	1.12	CA	1.02
SC	0.70	MI	-0.73	HI	-0.18
VA	-0.11	OH	-0.88	OR	-0.26
WV	0.11	WI	-0.38	WA	0.18

The root mean square in the analysis of variance for state within division, averaged across all divisions, is 1.72 percent. Recall that only in the unweighted Division 2 analysis were the differences between states significant, it must be emphasized that the estimates in Table 8 do not represent well-measured interstate differences. The fact that the estimated effects are substantial in magnitude but are still not statistically significant tells us that the power of these tests to find interstate differences, given the sample sizes of the PES, is not as great as might be desired.

Another approach to the power problem is to consider the effect of reducing the size of the census extract used in analysis of surrogate variables by a factor of 25, the ratio of the census extract to the PES sample sizes. If we divide by 25 each of the chi-square test statistics summarized in Table 3, then in only 27 out of 99 PSGs would

interstate differences have been significant for allocation rate (compared to 94 out of 99 PSGs with the full sample). Similarly, significant differences would have been found for 53 out of 99 PSGs for mail return rate (compared to 92 out of 99 PSGs with the full sample), and for 14 out of 84 for substitution rate (compared to 74 out of 84). Substitution rates are comparable in magnitude to undercount rates; after our hypothetical reduction of sample size, we obtain similar numbers of significant tests for substitution and undercount rates. It is plausible that with a much larger sample we would have found many more significant interstate differences, although one can only speculate on whether they would have been large enough to be of substantive concern.

5. DISCUSSION

This paper evaluates interstate heterogeneity in undercount rate and other census variables in the 1990 Census.

The evaluation used 1990 Census data and 1990 PES data. When this research was first embarked upon, the PES data were unavailable and were not expected to become available for analysis before the scheduled completion date. Surrogate variables from the 1990 Census were tested for significant heterogeneity among states within PSG. At the PSG level, state effect was significant ($\alpha = .05$) for 84%-95% of its PSGs for the various surrogate variables.

ANOVA on linearized undercount based on the PES data at the PSG level showed significant ($\alpha = .05$) state effects for 19 out of 99 PSGs. The significant results were concentrated in the PSGs in non-PMSA areas. Ten out of 32 such PSGs had significant state effects. This suggests that the poststratification in the relatively nonurban areas was not as successful as in the more urbanized areas.

How can we explain the different findings of the two studies? The two data sets had very different sample sizes, *i.e.*, the Census data had 125,000 block clusters but the PES data had 5,293 block clusters. It is therefore not surprising that small differences between states on surrogate variables would be statistically significant although corresponding differences would not be demonstrable with respect to undercount rates.

Furthermore, the correlations between the undercount rate and the surrogate variables are low as shown in Table 1. Therefore, any generalization from surrogate variables to undercount rates is somewhat conjectural. Given the modest correlation between undercount rates and surrogate variables, we prefer to give greater weight to the analysis of the PES data.

We conclude from these data that there are no demonstrable differences in average undercount rate between states within each division, after adjusting for PSG effects. While there is weak evidence for a difference between

New Jersey and New York within the Mid-Atlantic division, this result must be downweighted in the context of the number of divisions (nine) for which the test was performed. We conclude that if adjustment of population counts had been carried out based on the 1990 PES, no state would have been able to show that the poststratification was manifestly unfair in that it underadjusted that state relative to what direct state estimates showed that it deserved.

As the review in Section 2 shows, there is no consensus on whether or not between-state heterogeneity in undercount rates within PSG which is of substantial magnitude, although not large enough to be accurately measured by PES, would systematically affect the gain in accuracy obtained by synthetic adjustment. Nonetheless, the differences between states that were identified in analysis of the PES, together with the ancillary evidence of the surrogate variable analyses, make it appear likely that heterogeneity between states will again be an issue in coverage measurement for the year 2000 census, especially for the larger states for which these coverage differences can be most accurately measured. Fay and Thompson (1993) argue that a coverage measurement sample for 2000 should be designed to support direct (rather than synthetic) estimates of undercount for all states, although a CNSTAT panel (CNSTAT 1994) warns that for some states this could impose a highly inefficient sample allocation. Research over the intervening years must address the development of a combination of sample design and estimation methods that will produce defensible estimates of population by state.

ACKNOWLEDGEMENTS

The authors express their appreciation to the referees, an associate editor and editor for their comments which led to an improved version of this manuscript. This paper reports general results of research undertaken by the authors. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau or Harvard University. Zaslavsky's work was supported by Joint Statistical Agreements 90-23 and 91-31 and a contract between the Census Bureau and the National Opinion Research Center. The third author was with the Bureau of the Census when the research was in progress.

APPENDIX

Testing for Interstate Differences Using Linearized Statistics

A two-way ANOVA for adjustment factors in state parts yields an intuitively meaningful summary of the relative contributions of state and PSG effects to the variation in adjustment factors. Because the sampling unit of the PES is the block cluster rather than the state part,

these models do not yield valid statistical tests of the significance of the state effects.

Consider a statistic whose sample estimate for a state or state part is a weighted mean of the sample estimates in each component block or BP. Significance of the state effects for this statistic within a PSG could be evaluated by one-way ANOVA with the included block parts as units (corresponding to PSUs), or aggregated across PSGs by two-way ANOVA for state and PSG effects.

The sample adjustment factor estimate $(WCE/WE)/(WM/WP)$ is a nonlinear function of sample counts. In small primary sampling units (PSUs) such as block parts this nonlinearity may be very noticeable, especially when the number of matches in a PSU is very small or zero so that the sample estimate of the adjustment factor is large or infinite. In this situation, if PSU sample estimates are treated as data, the additive assumptions of ANOVA are violated. Useful tests may be recovered, however, by using a linearized version of the statistic of interest.

Suppose that we are interested in a parameter $Z = f(X)$ where X is a vector of population proportions in certain cells. Let \bar{x} , x_i represent the corresponding sample proportions in the entire sample and in PSU i respectively, so $\bar{x} = \sum N_i x_i / \sum N_i$ is a size-weighted average of block cell proportions. Let $f_1(X)$ be the gradient of f at X . Then by Taylor linearization $f(\bar{x}) - f(X) \approx f_1(X)'(\bar{x} - X) = \sum N_i f_1(X)'x_i / \sum N_i - f_1(X)'X$, i.e., we may treat the problem as one of inference regarding the quantities (pseudo-observations) $z_i = f_1(X)'x_i$. Because the approximate (linearized) influence of PSU i on the estimate $f(\bar{x})$, that is, the difference between the estimate with and without PSU i included, is $N_i f_1(X)'(x_i - \bar{x})$, we may describe this as a method based on influence statistics (Hampel *et al.* 1986) or the infinitesimal jackknife (Efron 1982, Chapter 6).

To derive a sensible variance model, suppose that we may regard PSU i as sample (not necessarily SRS) from a superpopulation with cell proportions X_i . A simple model is then, for some covariance matrices U_i and V_i ,

superpopulation model:

$$E(X_i) = X, \quad \text{Var}(X_i) = V_i$$

and

sampling model:

$$E(x_i | X_i) = X_i, \quad \text{Var}(x_i | X_i) = U_i.$$

The sampling covariance U_i will typically be proportional to N_i^{-1} . A plausible and mathematically convenient specification for V_i is $V_i \propto N_i^{-1}$ (i.e., smaller PSUs more variable than larger ones), so $\text{Var} z_i = \sigma^2/N_i$ for some constant σ^2 . The corresponding linear model weight for PSU i is N_i so the model-based estimate of the mean agrees with the design-based estimate obtained by aggregating the cell counts if N_i is a weighted size measure.

In the case of the adjustment factor $\hat{R} = (WCE/WE)/(WM/WP)$, the pseudo-observations are of the form $z_i = f_1(X)'(x_i - \bar{x}) =$

$$\hat{R} \left(\frac{WCE_i}{WCE} + \frac{WP_i}{WP} - \frac{WE_i}{WE} - \frac{WM_i}{WM} \right),$$

where WCE_i , WP_i , WE_i and WM_i are similar to the above for the i -th BP. We approximate the appropriate weight of a block part by $N_i = (WE_i + WP_i)/2$.

If the variance specifications of the model are inaccurate so there is some heteroscedasticity, or if the distribution is very long-tailed, then there will be a long-tailed distribution of residuals, making the tests at least slightly liberal. Some care must be taken to note the presence of outliers signaling this heteroscedasticity, for example, outlying blocks due to large-scale geocoding errors.

The assumption of approximately independent observations in ANOVA may be violated in two ways. First, the PSUs are not selected by SRS but rather by a geographical stratification somewhat finer than reflected in the post-stratification scheme. To the extent that this geographical stratification reduces the sampling variance of the state effect estimates, inferences under the independence model will be somewhat conservative. Second, there will be correlations between adjustment factors for different block parts from the same block (in multi-PSG models). These will tend to make inferences assuming independence somewhat liberal. On the balance, we regard the tests performed here as useful.

REFERENCES

- ALHO, J.M., MULRY, M.H., WURDEMAN, K., and KIM, J. (1991). Estimating heterogeneity in the probabilities of enumeration for dual-system estimation. *Journal of the American Statistical Association*, 88, 1130-1136.
- BUREAU OF THE CENSUS (1990). Sample Selection Procedures for Performing Evaluation Study P12. STSD 1990 Coverage Studies and Evaluation Memorandum Series No. N-1, Memorandum from D. Bateman to L. Iskow and M. Lynch, October 3, 1990.
- BUREAU OF THE CENSUS (1991). Request for Block Split Level Data for Performing PES Evaluation Project P12. STSD 1990 Coverage Studies and Evaluation Memorandum Series No. N-2, Memorandum from J. Thompson to A. Jackson, January 30, 1991.
- COMMITTEE ON NATIONAL STATISTICS, PANEL TO EVALUATE ALTERNATIVE CENSUS METHODS (1994). *Counting People in the Information Age*. Washington D.C.: National Academy Press.
- EFRON, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. Philadelphia: Society for Industrial and Applied Mathematics (SIAM).
- FAY, R.E., and THOMPSON, J.H. (1993). The 1990 Post Enumeration Survey Statistical Lessons, in *Hindsight. Proceedings of the 1993 Annual Research Conference*. U.S. Bureau of the Census, 71-91.
- FREEDMAN, D.A., and NAVIDI, W.C. (1992). Should we have adjusted the U.S. Census of 1980? *Survey Methodology*, 18, 3-24.
- FREEDMAN, D.A., PISANI, R., and PURVIS, R. (1978). *Statistics*. New York: Norton.
- FREEDMAN, D.A., and WACHTER, K.W. (1993). Heterogeneity and Census Adjustment for the Inter-Censal Base. Technical Report No. 381, Department of Statistics, University of California at Berkeley.
- HAMPEL, F.R., RONCHETTI, E.M., ROUSSEEUW, P.J., and STAHEL, W.A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. New York: John Wiley and Sons.
- HENGARTNER, N., and SPEED, T.P. (1993). Assessing between-block heterogeneity within the poststrata of the 1990 Post Enumeration Survey. *Journal of the American Statistical Association*, 88, 1119-1125.
- HOGAN, H. (1992). The 1990 Post Enumeration Survey: An overview. *American Statistician*, 46, 261-269.
- HOGAN, H. (1993). The 1990 Post Enumeration Survey: Operations and results. *Journal of the American Statistical Association*, 88, 1047-1057.
- ISAKI, C.T., SCHULTZ, L.K., DIFFENDAL, G.J., and HUANG, E.T. (1988). On estimating census undercount in small areas. *Journal of Official Statistics*, 4, 95-112.
- KIM, J. (1991). 1990 PES Evaluation Project P12: Evaluation of Synthetic Assumption. 1990 Coverage Studies and Evaluation Memorandum Series No. N-4, internal memorandum, U.S. Bureau of the Census.
- MULRY, M.H., and SPENCER, B.D. (1993). Accuracy of the 1990 Census and undercount adjustment. *Journal of the American Statistical Association*, 88, 1080-1091.
- SCHAFER, J.L. (1993). Comment on Hengartner, N and Speed, T.P.'s Assessing between-block heterogeneity within the poststrata of the 1990 Post Enumeration Survey. *Journal of the American Statistical Association*, 88, 1125-1127.
- SCHIRM, A.L., and PRESTON, S.H. (1987). Census undercount adjustment and quality of geographic population distributions. *Journal of the American Statistical Association*, 82, 965-978.
- WACHTER, K.W., and FREEDMAN, D.A. (1992). Measuring Local Homogeneity 1990 Census Data. Technical Report, Department of Statistics, University of California at Berkeley.
- WOLTER, K.M., and CAUSEY, B.D. (1991). Evaluation of procedures for improving population estimates for small areas. *Journal of the American Statistical Association*, 86, 278-284.

Markov Chain Designs for One-Per-Stratum Sampling

F. JAY BREIDT¹

ABSTRACT

Classical results in finite population sampling tell us that systematic sampling is the most efficient equal-probability one-per-stratum design for certain kinds of autocorrelated superpopulations, but stratified simple random sampling may be much better than systematic sampling if the superpopulation is a trend with uncorrelated errors. What if the superpopulation consists of a trend plus autocorrelated errors? Intuitively, some sort of “compromise” between the two designs might be better than either. Such compromise designs are constructed in this paper and are shown to be examples of Markov chain designs, a wide class of methods for one-per-stratum selection from a finite population. These designs include as special cases systematic sampling, balanced systematic sampling and stratified simple random sampling with one sampling unit per stratum. First and second-order inclusion probabilities are derived for Markov chain designs, yielding the Horvitz-Thompson estimator and its variance. Efficiency of the Horvitz-Thompson estimator is evaluated using superpopulation models. Numerical examples show that new designs considered here can be more efficient than standard designs for superpopulations consisting of trend plus autocorrelated errors. An example of the implementation of Markov chain designs for the 1992 National Resources Inventory in Alaska is given.

KEY WORDS: Balanced systematic sampling; National Resources Inventory; Systematic sampling.

1. INTRODUCTION

A stratified sampling design, in which a finite population is divided into non-overlapping strata and samples are drawn from each stratum, is a common and effective technique for reducing sampling error. In practice, stratified sampling designs with only one sampling unit per stratum are widely used. Examples include stratified simple random sampling and systematic sampling with its variants (e.g., Murthy and Rao 1988).

Systematic samples are susceptible to systematic errors. In large-scale spatial samples, for example, sources of systematic error could include roads, powerlines, irrigation systems, and so forth. A favorite example is the system of “section roads” in areas of the United States covered by the public land survey. This grid-based system is built up from square tracts of land called sections, each one mile on a side, which are often bounded by roads in midwestern agricultural regions. A systematic sampler with a one-mile sampling interval and an unlucky random start might conclude that Iowa is covered by gravel roads!

Systematic sampling does have the advantage of efficiency when the sampled population is positively autocorrelated, as is often the case in temporal and spatial sampling problems, since it forces observations to be as distant and hence as uncorrelated as possible.

Both autocorrelation and systematic error are of concern in the National Resources Inventory (NRI), an area sample of the nonfederal lands in the United States conducted every five years by the Soil Conservation Service of the

United States Department of Agriculture. NRI data items, collected by a combination of remote sensing and ground observation, include soil characteristics, land use, agricultural practices, erosion measures, and so on.

The 1992 NRI sample design for the northwestern region of the state of Alaska is a controlled version of one-per-stratum sampling. The region was divided into twenty-minute bands of latitude. Each band was divided into 500,000-acre strata. Each stratum was divided into a 10×10 grid of cells indexed by latitude and longitude, and one cell per stratum was selected. Selection moved from east to west across the strata within a particular twenty-minute band. The random numbers which determined the longitude cells of the selected units and the random numbers which determined the latitude cells evolved as two independent Markov chains. (Basic results on Markov chains used in this paper can be found in an introductory text on stochastic processes such as Taylor and Karlin 1984). Details of the design are given in Section 2.

How does this *ad hoc* design compare to more standard one-per-stratum designs? It turns out, as shown in Section 2, that simple Markov chain techniques can describe a broad class of equal-probability designs for one-per-stratum selection from a finite population. This class includes standard techniques such as stratified simple random sampling, systematic sampling and balanced systematic sampling, as well as the Alaska designs described above. It is also easy to generate new designs within this class. This unified treatment of one-per-stratum designs allows for comparisons of efficiency.

¹ F. Jay Breidt, Iowa State University, Department of Statistics, Ames, IA 50011-1210, U.S.A.

First and second-order inclusion probabilities for all of these designs are derived in Section 3, yielding the Horvitz-Thompson estimator and its variance. As in much of the relevant literature (Madow and Madow 1944; Cochran 1946; Sedransk 1969; Bellhouse and Rao 1975; Wolter 1985; Bellhouse 1988; *etc.*) the average design variance of the Horvitz-Thompson estimator is evaluated under a variety of superpopulation models. Compact expressions for model-averaged design variances are obtained. Numerical examples in Section 4 show that designs introduced in this paper can be more efficient than standard one-per-stratum designs for superpopulations consisting of trend plus autocorrelated errors. Discussion follows in Section 5.

Though our motivating example is two-dimensional, one-dimensional designs will be considered throughout. Most proofs and derivations are straightforward and are omitted for brevity.

2. MARKOV CHAIN DESIGNS

Consider the problem of sampling from a finite population of $N = na$ labeled units, denoted by

$$\begin{aligned} U &= \{1, \dots, N\} \\ &= \{1, \dots, a, a + 1, \dots, 2a, \dots, \\ &\quad (n - 1)a + 1, \dots, na\}. \end{aligned}$$

The value of a study variable $y_k = y_{(i-1)a+j} = y_{ij}$ is associated with each label k ; the notation y_k or y_{ij} will be used for both random variables and realizations of random variables.

Here n is the sample size and a is the *sampling interval*. The n subsets

$$\{(i - 1)a + 1, \dots, (i - 1)a + a\} \quad (i = 1, \dots, n)$$

will be referred to as *strata*. The goal is to select one unit per stratum. Often, a stratified sampling design is defined to be one in which independent probability samples are selected in each stratum, but the restriction to independence is not used here.

Given a doubly stochastic transition probability matrix P , a *Markov chain sample* is given by

$$s = \{R_1, a + R_2, \dots, (n - 1)a + R_n\},$$

where R_1, \dots, R_n is the Markov chain defined by P and $R_1 \sim \text{uniform}(1, \dots, a)$. Formally, then, a *Markov chain design* (MC) is a function $p(\cdot; P)$ such that

$$\begin{aligned} p(s; P) &= \Pr\{s = \{r_1, a + r_2, \dots, (n - 1)a + r_n\}\} \\ &= \Pr\{R_1 = r_1, R_2 = r_2, \dots, R_n = r_n\} \end{aligned}$$

$$= \begin{cases} P_{r_{n-1}, r_n} P_{r_{n-2}, r_{n-1}} \cdots P_{r_1, r_2} / a, \\ \quad \text{for } r_1, \dots, r_n \in \{1, \dots, a\}, \\ 0, \quad \text{otherwise.} \end{cases}$$

MC designs as defined in this paper are related to the designs given in Chandra, Sampath and Balasubramani (1992), in which a $1 \times N$ vector of initial selection probabilities and a $N \times N$ transition probability matrix of periodicity n determine a without-replacement sampling scheme. Chandra *et al.* focus on producing designs with strictly positive second-order inclusion probabilities. They do not explicitly consider the one-per-stratum designs of this paper, which can be imbedded in their structure in a straightforward way by constructing the appropriate initial probability vector and transition probability matrix.

The following result is useful in deriving the probabilistic features of MC designs.

Result 1 Consider a Markov chain for which the transition probability matrix P is doubly stochastic (*i.e.*, all row sums and all column sums equal one) and R_1 has a discrete uniform distribution, with mass $1/a$ on each of the states $1, \dots, a$. Then R_i has a discrete uniform distribution on the states $1, \dots, a$ for all i . In particular, R_i has mean $(a + 1)/2$ and variance $V(R_i) = (a^2 - 1)/12$.

Some special cases of MC designs are of interest.

Stratified simple random sampling. If the transition probability matrix is

$$H = [1/a]_{j,j'=1}^a,$$

then

$$\begin{aligned} \Pr\{R_{i'} = j' \mid R_i = j\} &= 1/a = \Pr\{R_{i'} = j'\} \\ (j, j' &= 1, \dots, a; i < i'), \end{aligned}$$

which, together with the Markov property, implies that R_1, \dots, R_n are probabilistically independent. In this case, the MC design is stratified simple random sampling with one unit per stratum (ST).

Systematic sampling. If the transition probability matrix is I , the $a \times a$ identity matrix, then

$$\Pr\{R_{i'} = j' \mid R_i = j\} = \begin{cases} 1, & j = j', \\ 0, & j \neq j', \end{cases}$$

so that R_1, \dots, R_n are deterministically related. Thus,

$$s = \{R_1, a + R_1, \dots, (n - 1)a + R_1\},$$

and so the MC design is systematic sampling (SY).

Compromise designs. Intuitively, ST and SY are at opposite “extremes” in some sense. If $\rho \in [0,1]$, then

$$G_\rho = \rho H + (1 - \rho)I$$

is doubly stochastic. If $\rho = 0$, the design is SY and if $\rho = 1$, the design is ST. Any other choice of ρ will yield a sequence consisting of “runs” of SY samples. Thus, the class G_ρ includes ST and SY, as well as a continuum of “compromise” MC designs.

Other convex combinations of doubly stochastic matrices could be considered. The class of doubly stochastic matrices is also closed under matrix multiplication, transposition, and row and column permutation, so there are many ways to create MC designs.

Balanced systematic sampling. Murthy (1967, §5.9d) describes a one-per-stratum selection method which he calls *balanced systematic sampling* (BA). This method gives samples

$$s = \{R_1, a + (a + 1 - R_1), \dots, (n - 2)a + R_1, \\ (n - 1)a + (a + 1 - R_1)\}$$

for n even and

$$s = \{R_1, a + (a + 1 - R_1), \dots, (n - 2)a + \\ (a + 1 - R_1), (n - 1)a + R_1\}$$

for n odd. An interesting feature of this design is that if n is even and the population is perfectly linear ($y_{ij} = \beta_0 + \beta_1[(i - 1)a + j]$), then the sample mean equals the population mean for any sample. With the transition probability matrix,

$$J = \begin{pmatrix} 0 & 0 & \dots & 0 & 1 \\ 0 & 0 & \dots & 1 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 1 & \dots & 0 & 0 \\ 1 & 0 & \dots & 0 & 0 \end{pmatrix}_{a \times a},$$

BA is a MC design.

Alaska NRI design. As described in Section 1, the 1992 NRI sample design for the northwestern region of the state of Alaska used two independent Markov chains in the controlled selection of latitude and longitude cells. The transition probability matrix for longitude cells, P_{long} , is given in Table 1. This design, henceforth denoted AK, is a MC design since P_{long} is doubly stochastic. Most of the transition probabilities are close to 0.10, so most “step sizes”

are approximately equally likely. Note, however, that mass has been removed from on and near the back diagonal and placed in the upper left and lower right corners, so that P_{long} discourages large east to west steps, such as from cell one to cell ten, and discourages small steps, such as from cell ten to cell one. On the other hand, P_{long} encourages steps of around length ten, such as from cell two to cell one, two or three. The realized sample of longitude cells is thus well-dispersed east to west, like a systematic sample would be, but its additional randomness guards against systematic error. Similarly, the Markov chain for latitude cells was set up to give good spatial dispersion north to south.

Table 1

Transition probability matrix for Markov chain sample of longitude cells, 1992 National Resources Inventory, Alaska. Entries are the conditional probabilities of selecting cell j' of stratum $i + 1$ given that cell j of stratum i was selected.

Cell j of stratum i	Cell j' of stratum $i + 1$									
	1	2	3	4	5	6	7	8	9	10
1	0.05	0.15	0.15	0.15	0.15	0.15	0.10	0.10	0	0
2	0.15	0.15	0.15	0.10	0.10	0.10	0.10	0.10	0.05	0
3	0.15	0.15	0.10	0.10	0.10	0.10	0.05	0.05	0.10	0.10
4	0.15	0.10	0.10	0.10	0.10	0.10	0.10	0.05	0.10	0.10
5	0.15	0.10	0.10	0.10	0.05	0.05	0.10	0.10	0.10	0.15
6	0.15	0.10	0.10	0.10	0.05	0.05	0.10	0.10	0.10	0.15
7	0.10	0.10	0.05	0.10	0.10	0.10	0.10	0.10	0.10	0.15
8	0.10	0.10	0.05	0.05	0.10	0.10	0.10	0.10	0.15	0.15
9	0	0.05	0.10	0.10	0.10	0.10	0.10	0.15	0.15	0.15
10	0	0	0.10	0.10	0.15	0.15	0.15	0.15	0.15	0.05

3. HORVITZ-THOMPSON ESTIMATION UNDER MC

Write the population total as

$$t = \sum_U y_k = \sum_{i=1}^n \sum_{j=1}^a y_{(i-1)a+j} = \sum_{i=1}^n \sum_{j=1}^a y_{ij}.$$

For all k , the first-order inclusion probabilities of a MC design are given by

$$\pi_k = \Pr\{k \in s\} = \Pr\{R_i = j\} = 1/a$$

and for $k \leq l$, the second-order inclusion probabilities are given by

$$\pi_{kl} = \begin{cases} 1/a, & \text{for } i = i', j = j', \\ 0, & \text{for } i = i', j \neq j', \\ P_{jj'}^{(i'-i)}/a, & \text{for } i < i'. \end{cases}$$

The design-unbiased Horvitz-Thompson estimator (Horvitz and Thompson 1952) for the population total is then

$$\hat{t}_\pi = \sum_s y_k / \pi_k = \sum_{i=1}^n \frac{y_i R_i}{1/a} = a \sum_{i=1}^n \sum_{j=1}^a y_{ij} I_{\{R_i=j\}},$$

where

$$I_{\{R_i=j\}} = \begin{cases} 1, & \text{if } R_i = j, \\ 0, & \text{if } R_i \neq j. \end{cases}$$

The design covariances of the indicators $I_{\{R_i=j\}}$ are given by

$$\begin{aligned} C_{MC}(I_{\{R_i=j\}}, I_{\{R_{i'}=j'\}}) &= E_{MC}[I_{\{R_i=j\}} I_{\{R_{i'}=j'\}}] - \\ &E_{MC}[I_{\{R_i=j\}}] E_{MC}[I_{\{R_{i'}=j'\}}] \\ &= \pi_{(i-1)a+j, (i'-1)a+j'} - \\ &\pi_{(i-1)a+j} \pi_{(i'-1)a+j'}, \end{aligned}$$

and so the design variance of \hat{t}_π is

$$\begin{aligned} V_{MC}(\hat{t}_\pi) &= a^2 \sum_{i=1}^n \sum_{j=1}^a \left(\frac{1}{a} - \frac{1}{a^2} \right) y_{ij} y_{ij} \quad (1) \\ &+ a^2 \sum_{i=1}^n \sum_{j=1}^a \sum_{j' \neq j} \left[0 - \frac{1}{a^2} \right] y_{ij} y_{ij'} \\ &+ 2a^2 \sum_{i=1}^n \sum_{i' > i} \sum_{j=1}^a \left[\frac{P_{jj'}^{(i'-i)}}{a} - \frac{1}{a^2} \right] y_{ij} y_{i'j'} \\ &+ 2a^2 \sum_{i=1}^n \sum_{i' > i} \sum_{j=1}^a \sum_{j' \neq j} \left[\frac{P_{jj'}^{(i'-i)}}{a} - \frac{1}{a^2} \right] y_{ij} y_{i'j'}. \end{aligned}$$

Since the design variance depends on all the values of the study variable in the finite population, (1) is not easily used for comparing designs. Following Cochran (1946), assume that the values of the study variable are generated from the superpopulation model

$$\xi : y_{ij} = \mu_{ij} + e_{ij},$$

where the μ_{ij} are fixed and the e_{ij} are random variables with $E_\xi[e_{ij}] = 0$, $V_\xi(e_{ij}) = \sigma_{ij}^2$ and $C_\xi(e_{ij}, e_{i'j'}) = \sigma_{ij, i'j'}$. Then designs can be compared on the basis of model-averaged design variance.

Proposition 1 Under the superpopulation model ξ , the average design variance of the Horvitz-Thompson estimator is

$$\begin{aligned} E_\xi[V_{MC}(\hat{t}_\pi)] &= a^2 V_{MC} \left[\sum_{i=1}^n \mu_{iR_i} \right] + \\ &(a-1) \sum_{i=1}^n \sum_{j=1}^a \sigma_{ij}^2 - \sum_{i=1}^n \sum_{j=1}^a \sum_{j' \neq j} \sigma_{ij, i'j'} \\ &+ 2a \sum_{i=1}^n \sum_{i' > i} \sum_{j=1}^a \sum_{j'=1}^a \sigma_{ij, i'j'} \left[\frac{P_{jj'}^{(i'-i)}}{a} - \frac{1}{a} \right] \end{aligned}$$

for any MC design. Note that if μ_{ij} is independent of j , then $V_{MC}[\sum_{i=1}^n \mu_{iR_i}] = 0$.

The following proposition gives a sufficient condition under which no MC design has worse average design variance than SY.

Proposition 2 Consider an uncorrelated additive model,

$$\xi : y_{ij} = \mu_{ij} + e_{ij} = \alpha_i + \beta_j + e_{ij},$$

where $E_\xi[e_{ij}] = 0$, $V_\xi(e_{ij}) = \sigma_{ij}^2$ and $C_\xi(e_{ij}, e_{i'j'}) = 0$. Then

$$E_\xi[V_{SY}(\hat{t}_\pi)] \geq E_\xi[V_{MC}(\hat{t}_\pi)]$$

for all MC designs.

Proof From Proposition 1, the only term of interest is $V_{MC}[\sum_{i=1}^n \mu_{iR_i}]$, which under SY is

$$V_{SY} \left[\sum_{i=1}^n \mu_{iR_i} \right] = V_{SY} \left[\sum_{i=1}^n \alpha_i + n\beta_{R_1} \right] = n^2 V(\beta_{R_1}),$$

while under a general MC design,

$$V_{MC} \left[\sum_{i=1}^n \mu_{iR_i} \right] = \sum_{i=1}^n \sum_{i'=1}^n C_{MC}(\beta_{R_i}, \beta_{R_{i'}}).$$

Since $C_{MC}(\beta_{R_i}, \beta_{R_{i'}}) \leq V(\beta_{R_1})$, the proposition follows. \square

Some specific models are considered in the next five subsections.

3.1 Random Permutation Model

A model for a population in random order is a permutation model, in which a realization of the measurements y_1, \dots, y_N is given by one of the $N!$ equally likely permutations of N fixed values. This model can be written as

$$\xi_1 : y_{ij} = \bar{y}_U + e_{ij},$$

where $\bar{y}_U = \sum_U y_k / N$. See Rao (1975) for more details. The following result is then a consequence of Theorem 2.1 of Rao and Bellhouse (1978).

Result 2 Under the random permutation model,

$$E_{\xi_1}[V_{MC}(\hat{t}_\pi)] = (N^2/n)(1 - n/N) \sum_U (y_k - \bar{y}_U)^2 / (N - 1)$$

for any MC design.

Thus, the average variance over all permutations is exactly $V_{SI}(\hat{t}_\pi)$, where SI denotes (unstratified) simple random sampling without replacement. For SY, this result is originally due to Madow and Madow (1944). See also Sedransk (1969).

3.2 Stratification Effects Model

A model for a population with stratification effects is

$$\xi_2: y_{ij} = \alpha_i + e_{ij},$$

where the α_i are fixed constants and e_{ij} are uncorrelated random variables with mean zero and variance σ^2 . Note that if $\alpha_i \equiv \mu$, then ξ_2 is an alternative to ξ_1 as a model for a population in random order.

Result 3 Under the stratification effects model,

$$E_{\xi_2}[V_{MC}(\hat{t}_\pi)] = na(a - 1)\sigma^2$$

for any MC design.

3.3 Linear Trend Model

A model for a population with a linear trend is

$$\xi_3: y_{ij} = \beta_0 + \beta_1[(i - 1)a + j] + e_{ij},$$

where β_0 and β_1 are fixed constants and e_{ij} are uncorrelated $(0, \sigma^2)$ random variables.

Result 4 Under the linear trend model ξ_3 ,

$$E_{\xi_3}[V_{MC}(\hat{t}_\pi)] = \beta_1^2 a^2 V_{MC}\left[\sum_{i=1}^n R_i\right] + na(a - 1)\sigma^2 \quad (2)$$

for any MC design. Since ξ_3 is additive, no MC design has a larger expected variance under a linear trend model than SY.

The only design-dependent term in (2) is $V_{MC}[\sum_{i=1}^n R_i]$. Under SY, $\sum_{i=1}^n R_i = nR_1$, so that

$$V_{SY}\left[\sum_{i=1}^n R_i\right] = n^2 V(R_1),$$

while under ST,

$$V_{ST}\left[\sum_{i=1}^n R_i\right] = nV(R_1).$$

Under BA, for n even,

$$V_{BA}\left[\sum_{i=1}^n R_i\right] = V_{BA}\left[\frac{n}{2}R_1 + \frac{n}{2}(a + 1 - R_1)\right] = 0.$$

This implies that if the population is perfectly linear ($\sigma^2 = 0$), then

$$E_{\xi_3}[V_{BA}(\hat{t}_\pi)] = 0,$$

so that $\hat{t}_\pi = t$ for all samples, as noted by Murthy (1967, p. 165).

Result 5 Under the linear trend model ξ_3 ,

$$\begin{aligned} E_{\xi_3}[V_{BA}(\hat{t}_\pi)] &\leq E_{\xi_3}[V_{ST}(\hat{t}_\pi)] \\ &\leq E_{\xi_3}[V_{G_\rho}(\hat{t}_\pi)] \\ &\leq E_{\xi_3}[V_{SY}(\hat{t}_\pi)] = \max_{MC} E_{\xi_3}[V_{MC}(\hat{t}_\pi)], \end{aligned} \quad (3)$$

where the middle term is monotone increasing with decreasing $\rho \in [0, 1]$. If n is even, the left-hand side of (3) equals $\min_{MC} E_{\xi_3}[V_{MC}(\hat{t}_\pi)]$.

3.4 Periodic Population Model

A simple model for a population showing a deterministic periodicity with period p is the sine wave model

$$\xi_4: y_{ij} = \alpha \sin\left\{\frac{2\pi}{p}[(i - 1)a + j]\right\} + e_{ij},$$

where e_{ij} are uncorrelated random variables with mean zero and variance σ^2 .

Result 6 Under the periodic population model ξ_4 ,

$$\begin{aligned} E_{\xi_4}[V_{MC}(\hat{t}_\pi)] &= a^2 \alpha^2 V_{MC}\left[\sum_{i=1}^n \sin\frac{2\pi}{p}[(i - 1)a + R_i]\right] \\ &\quad + na(a - 1)\sigma^2 \end{aligned}$$

for any MC design.

Denote the sine wave model ξ_4 with $p = a$ by ξ_{4a} . Under ξ_{4a} ,

$$\sin\left\{\frac{2\pi}{p}[(i - 1)a + j]\right\} = \sin\frac{2\pi j}{a},$$

so that the model is additive and no MC design has larger expected design variance under ξ_{4a} than SY, highlighting the fact that SY is inappropriate for a population containing a periodicity with period equal to the sampling interval (Madow and Madow 1944). This result generalizes as follows.

Result 7 If $\mu_{ij} \equiv \beta_j$ in ξ , then ξ is a model for a population showing a deterministic periodicity with period equal to the sampling interval, a . The model ξ is additive and so no MC design has larger expected design variance under ξ than SY.

3.5 Autocorrelated Model

Beginning with Cochran (1946), many authors have compared ST, SY and simple random sampling under an autocorrelated superpopulation model. See Bellhouse (1988, §4) for a review.

Consider the following autocorrelation model due to Cochran (1946):

$$\xi_5: y_{ij} = \mu + e_{ij}$$

where $\sigma_{ij,i'j'} = \gamma[(i' - i)a + j' - j]$ for $i' \geq i$.

Result 8 Under the autocorrelated model ξ_5 ,

$$E_{\xi_5}[V_{MC}(\hat{t}_\pi)] = na(a-1)\gamma(0) - 2n \sum_{h=1}^{a-1} \gamma(h)(a-h) + 2a \sum_{h=1}^{n-1} \sum_{j=1}^a \sum_{j'=1}^a \gamma(ha + j' - j)(n-h) \left(P_{jj'}^{(h)} - \frac{1}{a} \right)$$

for any MC design.

Result 9 If, for $h \geq 0$, $\gamma(h)$ is non-negative, non-increasing and convex, i.e.,

$$\gamma(h) \geq 0, \gamma(h) \geq \gamma(h+1) \quad \text{and}$$

$$\gamma(h+2) - 2\gamma(h+1) + \gamma(h) \geq 0,$$

then $E_{\xi_5}[V_{SY}(\hat{t}_\pi)] = \min_{MC} E_{\xi_5}[V_{MC}(\hat{t}_\pi)]$.

This result is a corollary of a theorem due to Hájek (1959), given as Theorem 4.1 of Bellhouse (1988); Bellhouse clarified the conditions under which the theorem holds. Hájek's theorem generalized an earlier result due to Cochran (1946), who compared SY, ST and simple random sampling.

4. EFFICIENCY: SOME NUMERICAL EXAMPLES

An important class of models for time series and spatial processes consists of a low-order polynomial trend plus an autocorrelated error sequence. A simple example is

$$\xi_{(\beta,\phi)}: y_{ij} = \beta_0 + \beta_1[(i-1)a + j] + e_{ij},$$

where the autocorrelation structure is that of a first-order autoregressive (AR) model,

$$\sigma_{ij,i'j'} = \gamma[(i' - i)a + j' - j] = \sigma^2 \phi^{(i' - i)a + j' - j}$$

for $i' \geq i$ and $|\phi| < 1$. The average design variance under this model is obtained from Results 4 and 8. For different choices of β_1 and ϕ , the ratio of expected design variances,

$$E_{\xi}[V_{MC}(\hat{t}_\pi)]/E_{\xi}[V_{SY}(\hat{t}_\pi)], \quad (4)$$

is given in Table 2 for various MC designs. Also tabled is the optimal G_ρ design, obtained by minimizing (4) with respect to ρ . Use of this design is only feasible if superpopulation parameters are known, so it is tabled merely as a benchmark and not as a competitor.

When $\beta_1 \neq 0$ and $\phi = 0$, the model is ξ_3 and the tabled values agree with Result 5: SY is the worst MC design and BA is the best, with $G_{1/3}$, $G_{2/3}$ and ST falling between them. Though BA does extremely well for this model, any non-SY MC design would be a good choice.

When $\beta_1 = 0$ and $\phi \neq 0$, $\xi_{(\beta,\phi)}$ is a special case of model ξ_5 . For $\phi > 0$, Result 9 and the table agree that SY is most efficient since it makes the sample as "spread out" as possible, but for weak autocorrelation, the other MC designs are competitive. BA is very poor for this model, because the design ensures that every other pair R_i, R_{i+1} will be no more than a units apart. (For the same reason, BA is good for a negatively autocorrelated population.) AK, $G_{1/3}$ and $G_{2/3}$ outperform ST, because each of these designs encourages state transitions of around length a .

Similar results are obtained for the superpopulation model

$$\xi_{(\alpha,\phi)}: y_{ij} = \alpha \sin \frac{2\pi j}{a} + e_{ij},$$

where $\sigma_{ij,i'j'}$ is as above. Table 2 gives the ratio of expected design variances (4) under this model, obtained from Results 6 and 8.

When $\alpha \neq 0$ and $\phi = 0$, the model is ξ_{4a} and SY performs badly, as indicated by Result 7. Even for $\phi \neq 0$, SY performs well only when the periodicity is swamped by highly-correlated noise.

Note that no design dominates Table 2: each of SY, $G_{1/3}$, $G_{2/3}$, ST, BA and AK is the best at least once among those considered. For a moderate trend and high autocorrelation, AK, $G_{1/3}$ and $G_{2/3}$ can beat standard MC designs. Overall, Table 2 suggests that some non-standard MC designs, such as $G_{2/3}$ and AK, do reasonably well for a variety of populations: retaining much of the efficiency of SY against an autocorrelated population, while still guarding against systematic effects in other kinds of populations.

Table 2

Ratio of expected design variance under MC to expected design variance under SY for superpopulation consisting of trend (line with slope β_1 or sine wave with period a and amplitude α) plus autoregressive (AR) errors ($N = 1,000$, $\sigma^2 = 100$, $a = 10$). Here G_{ρ^*} is the optimal compromise design, where ρ^* is a function of superpopulation parameters. Ratio for the best realizable design in each row (if not SY) is italicized.

Model	ϕ	Markov Chain Design					
		$G_{1/3}$	$G_{2/3}$	ST	BA	AK	G_{ρ^*} (ρ^*)
Line + AR $\beta_1 = 0.7$	-0.5	0.2322	0.2085	0.2001	<i>0.1666</i>	0.2056	0.2001 (1.0000)
	0.0	0.2220	0.1983	0.1903	<i>0.1821</i>	0.1957	0.1903 (1.0000)
	0.1	0.2187	0.1950	0.1871	<i>0.1825</i>	0.1921	0.1871 (1.0000)
	0.5	0.1922	0.1702	<i>0.1645</i>	0.1754	0.1659	0.1645 (1.0000)
	0.9	0.0980	0.0778	<i>0.0742</i>	0.0768	<i>0.0762</i>	0.0742 (1.0000)
Line + AR $\beta_1 = 0.4$	-0.5	0.4504	0.4328	0.4262	<i>0.3647</i>	0.4304	0.4262 (1.0000)
	0.0	0.4344	0.4172	0.4114	<i>0.4054</i>	0.4153	0.4114 (1.0000)
	0.1	0.4291	0.4121	<i>0.4065</i>	0.4085	0.4094	0.4065 (1.0000)
	0.5	0.3853	0.3727	0.3724	0.4116	<i>0.3667</i>	0.3719 (0.8320)
	0.9	0.1876	<i>0.1835</i>	0.1914	0.2170	0.1848	0.1821 (0.5223)
Line + AR $\beta_1 = 0.1$	-0.5	0.9233	0.9190	0.9163	<i>0.7941</i>	0.9175	0.9163 (1.0000)
	0.0	0.9201	0.9177	0.9169	<i>0.9160</i>	0.9174	0.9169 (1.0000)
	0.1	0.9191	0.9175	0.9175	0.9349	<i>0.9156</i>	0.9174 (0.8156)
	0.5	<i>0.9160</i>	0.9289	0.9439	1.0606	0.9185	0.9135 (0.1997)
	0.9	<i>0.8621</i>	0.9787	1.0725	1.2710	1.0017	0.7888 (0.0981)
Pure AR	-0.5	0.9978	0.9956	0.9935	<i>0.8617</i>	0.9942	0.9935 (1.0000)
	0.0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000 (---)
	0.1	1.0009	1.0019	1.0028	1.0228	1.0001	1.0000 (0.0000)
	0.5	1.0179	1.0357	1.0536	1.1852	1.0245	1.0000 (0.0000)
	0.9	1.2517	1.4380	1.5814	1.8798	1.4734	1.0000 (0.0000)
Sine + AR $\alpha = 0.1$	-0.5	0.9929	0.9906	0.9884	<i>0.8578</i>	0.9892	0.9884 (1.0000)
	0.0	0.9947	0.9946	<i>0.9945</i>	0.9950	0.9946	0.9945 (1.0000)
	0.1	0.9955	0.9963	0.9972	1.0175	<i>0.9945</i>	0.9954 (0.1925)
	0.5	1.0110	1.0285	1.0462	1.1775	1.0173	0.9977 (0.0364)
	0.9	1.2178	1.3980	1.5371	1.8294	1.4322	0.9999 (0.0018)
Sine + AR $\alpha = 1.0$	-0.5	0.6747	0.6634	0.6586	<i>0.6008</i>	0.6604	0.6586 (1.0000)
	0.0	0.6603	0.6499	<i>0.6464</i>	0.6770	0.6477	0.6464 (1.0000)
	0.1	0.6554	0.6455	0.6425	0.6863	<i>0.6421</i>	0.6425 (1.0000)
	0.5	0.6149	0.6133	0.6196	0.7320	<i>0.6041</i>	0.6121 (0.5079)
	0.9	<i>0.3570</i>	0.3832	0.4126	0.5527	0.3877	0.3560 (0.2852)
Sine + AR $\alpha = 10.0$	-0.5	0.0668	0.0384	<i>0.0287</i>	0.1101	0.0323	0.0287 (1.0000)
	0.0	0.0656	0.0372	<i>0.0275</i>	0.1115	0.0311	0.0275 (1.0000)
	0.1	0.0652	0.0368	<i>0.0271</i>	0.1115	0.0307	0.0271 (1.0000)
	0.5	0.0622	0.0339	<i>0.0245</i>	0.1106	0.0277	0.0245 (1.0000)
	0.9	0.0529	0.0247	<i>0.0154</i>	0.1016	0.0187	0.0154 (1.0000)

5. DISCUSSION

The class of Markov chain designs has been defined and shown to include systematic sampling, stratified simple random sampling and balanced systematic sampling as special cases. Some new designs have been introduced (G_{ρ} , AK) and shown to be competitive with standard one-per-stratum designs under a variety of superpopulation models. In particular, the new designs work well in numerical examples for trending superpopulations with autocorrelated errors. This is the kind of population of concern in many area sampling problems, such as the 1992 National Resources Inventory in Alaska. A two-dimensional MC design implemented for that survey shows that one-dimensional MC designs might be usefully extended to a spatial sampling context, though further work on this extension is necessary.

Further work on variance estimation for MC designs is also needed. Because these are one-per-stratum designs, design-unbiased estimation of the variance of the Horvitz-Thompson estimator is not possible. The problem of variance estimation for one-per-stratum designs, particularly for SY, has received much attention. For example, Wolter (1985) discusses in detail eight different biased variance estimators for SY and evaluates their biases under superpopulation models. Work in this direction for the collapsed strata variance estimator (e.g., Cochran 1977, p. 139) under general MC designs is in progress.

ACKNOWLEDGEMENTS

This work was supported by the U.S. Department of Agriculture's Soil Conservation Service under SCS Cooperative Agreement No. 68-3A75-2-64. Jeff Goebel (Soil Conservation Service) and Wayne Fuller (Iowa State University) developed the MC design for the state of Alaska. The author thanks the anonymous referees for their constructive comments on an earlier version of this paper.

REFERENCES

- BELLHOUSE, D.R. (1988). Systematic sampling. In *Handbook of Statistics*. (Eds. P.R. Krishnaiah and C.R. Rao), Vol. 6. Amsterdam: North-Holland, 125-145.
- BELLHOUSE, D.R., and RAO, J.N.K. (1975). Systematic sampling in the presence of a trend. *Biometrika*, 62, 694-697.
- CHANDRA, K.S., SAMPATH, S., and BALASUBRAMANI, G.K. (1992). Markov sampling for finite populations. *Biometrika*, 79, 210-213.
- COCHRAN, W.G. (1946). Relative accuracy of systematic and stratified random samples for a certain class of populations. *Annals of Mathematical Statistics*, 17, 164-177.
- COCHRAN, W.G. (1977). *Sampling Techniques*, 3rd Ed. New York: Wiley.
- HÁJEK, J. (1959). Optimum strategy and other problems in probability sampling. *Časopis Pro Pěstování Matematiky*, 84, 387-423.
- HORVITZ, D.G., and THOMPSON, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.
- MADOW, W.G., and MADOW, L.H. (1944). On the theory of systematic sampling, I. *Annals of Mathematical Statistics*, 15, 1-24.
- MURTHY, M.N. (1967). *Sampling Theory and Methods*. Calcutta: Statistical Publishing Society.
- MURTHY, M.N., and RAO, T.J. (1988). Systematic sampling with illustrative examples. In *Handbook of Statistics*. (Eds. P.R. Krishnaiah and C.R. Rao), (Vol. 6). Amsterdam: North-Holland, 147-185.

- RAO, J.N.K. (1975). On the foundations of survey sampling. In *A Survey of Statistical Design and Linear Models*. (Ed. J.N. Srivastava). Amsterdam: North-Holland, 489-505.
- RAO, J.N.K., and BELLHOUSE, D.R. (1978). Optimal estimation of a finite population mean under generalized random permutation models. *Journal of Statistical Planning and Inference*, 2, 125-141.
- SEDRANSK, J. (1969). Some elementary properties of systematic sampling. *Skandinavisk Aktuarietidskrift*, 1-2, 39-47.
- TAYLOR, H.M., and KARLIN, S. (1984). *An Introduction to Stochastic Modeling*. Orlando: Academic Press.
- WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

Median Estimation Using Auxiliary Information

GLEN MEEDEN¹

ABSTRACT

The problem of estimating the median of a finite population when an auxiliary variable is present is considered. Point and interval estimators based on a non-informative Bayesian approach are proposed. The point estimator is compared to other possible estimators and is seen to perform well in a variety of situations.

KEY WORDS: Sample survey; Estimation; Median; Auxiliary variable; Quantile; Non-informative Bayes.

1. INTRODUCTION

The problem of estimating a population mean in the presence of an auxiliary variable has been widely discussed in the finite population sampling literature. The ratio estimator has often been used in such situations. For the problem of estimating a population median the situation is quite different. Only recently has this problem been discussed. Chambers and Dunstan (1986) proposed a method for estimating the population distribution function and the associated quantiles. They assumed that the value of the auxiliary variable was known for every unit in the population and their estimator came from a model-based approach. Rao *et al.* (1990) proposed ratio and difference estimators for the median using a design-based approach. Kuk and Mak (1989) proposed two other estimators for the population median. To use the Kuk and Mak estimators one only needs to know the values of the auxiliary variable for the units in the sample and its median for the whole population. The efficiencies of these estimators depend directly on the probability of 'concordance' rather than on the validity of an assumption of linearity between the variable of interest and the auxiliary variable.

Recently Meeden and Vardeman (1991) discussed a non-informative Bayesian approach to finite population sampling. This new approach uses the 'Polya posterior' as a predictive distribution for the unobserved members of the population once the sample has been observed. Often it yields point and interval estimates that are very similar to those of standard frequentist theory. Moreover it can be easy to implement in problems that are difficult for standard theory. In this note we show how this method can be used for the problem of estimating a population median when an auxiliary variable is present and compare it to some of the other proposed methods.

2. ESTIMATING THE MEDIAN

Consider a finite population containing N units. For the unit with label i let y_i denote the characteristic of interest and x_i the auxiliary variable. We assume that both y_i and x_i are real numbers and each is known for every unit in the population. Let s denote a typical sample of size n which was chosen by simple random sampling without replacement. We assume simple random sample for convenience, since in many problems of this type the sampling will often be more purposeful. Before considering the problem of estimating the median of the population we review some well known facts about the problem of estimating the mean.

Consider the super population model where it is assumed that for each i , $y_i = bx_i + u_ie_i$. Here b is an unknown parameter while the u_i 's are known constants and the e_i 's are independent identically distributed random variables with zero expectations. Since the population mean can be written as $N^{-1}(\sum_{i \in s} y_i + \sum_{j \notin s} y_j)$ we would expect $N^{-1}(\sum_{i \in s} y_i + \hat{b} \sum_{j \notin s} x_j)$ to be a sensible estimate of the mean whenever \hat{b} is a sensible estimate of b . One particular choice of \hat{b} is the weighted least squares estimator where the weights are determined by the u_i 's. For example if for all i , $u_i = \sqrt{x_i}$, the resulting estimator is just the usual ratio estimator. While if for all i , $u_i = x_i$, then $\hat{b} = n^{-1} \sum_{i \in s} (y_i/x_i)$ and the resulting estimator is one that was discussed by Basu (1971). (See also Royall (1970).) Using this super population setup it is easy to generate populations where the ratio estimator has smaller mean squared error than the Basu estimator and vice versa. A somewhat limited simulation study on a variety of populations found that the performance of the Basu estimator is quite similar to the performance of the ratio estimator although in the majority of the cases the ratio estimator performs better than the Basu estimator. This is not unexpected, given the wide use of the ratio estimator.

¹ Glen Meeden, School of Statistics, University of Minnesota, Minneapolis, MN 55455, U.S.A.

In Meeden and Vardeman (1991) a non-informative Bayesian approach to finite population sampling, based on the Polya posterior, was developed. For the simple problem where no auxiliary variable is present, given the observed values in the sample, it introduces a Polya urn distribution as a pseudo posterior distribution over the unobserved members of the population. This pseudo posterior distribution can be used to obtain point and interval estimates of a variety of population quantities. It is related to the Bayesian bootstrap of Rubin (1981) and the Dirichlet process prior of Ferguson (1973). When estimating the median it yields results similar to those of Binder (1982). A theoretical justification for it is a stepwise Bayes argument which yields the admissibility of the resulting estimators. See for example Meeden and Ghosh (1983). There the admissibility of the Basu estimator was demonstrated. In that case the Basu estimator was shown to arise from a 'posterior' which treats the known and unknown ratios, $r_i = y_i/x_i$ as exchangeable. Note that this is very similar in spirit to the super population model justification for this estimator given above, where the ratios $r_i = y_i/x_i$ were independent and identically distributed. We shall see that the stepwise Bayes logic underlying the Basu estimator for the mean carries over in a straight forward way to point and interval estimators for the median. Unfortunately this is not the case for some of the other estimators. One natural but perhaps naïve estimator which mimics in some sense the ratio estimator of the mean is just the ratio of the median of the y values in the sample to the median of the x values in the sample multiplied by the median of the x values in the population. There is no known model based theory which underlies this estimator as is the case for the ratio estimator of the mean.

In the Bayesian approach to finite population sampling one needs to specify a prior distribution. Then given a sample, inferences are based on the posterior distribution, which is the predictive distribution for the unseen members of the population given the units in the sample. In the stepwise Bayes approach, given the sample one always has a 'posterior' distribution but it does not arise from a single prior distribution. However this 'posterior' distribution can be used in the usual Bayesian manner to find point and interval estimators of parameters of interest. We now will show how the stepwise Bayes model which yields Basu's estimator for the mean can also be used when estimating the median. In this setup, given a sample, the predictive distribution for the unobserved ratios treats the observed and unobserved ratios as 'exchangeable'.

For definiteness suppose our sample contains the first n units of the population. We construct an urn which contains n balls where ball i is given the value of the i -th observed ratio, say r_i . We begin by selecting a ball at random from the urn and the observed value is assigned to the unobserved unit $n + 1$. This ball and an additional ball with the same value is returned to the urn. Another

ball is chosen from the urn and its value is assigned to the unobserved unit $n + 2$. This ball and another with the same value are returned to the urn. This process is continued until all of the unobserved units have been assigned a ratio. Once they have all been assigned a value we have observed one realization from our 'posterior' distribution for the unseen ratios given the sample of seen ratios. If in this process the unobserved unit j has been assigned the ratio with value r we then assign its y_j value to be rx_j . Hence using simple Polya sampling we have created a predictive distribution for the unobserved units given the sample. We call this predictive distribution the 'Polya posterior'. It is easy to check that this predictive distribution gives the Basu estimator when estimating the population mean under squared error loss.

Given the sample the 'Polya posterior' yields a predictive distribution for the unobserved members of the population and hence a predictive distribution for the median as well. From the decision theory point of view the usual loss function is absolute error when estimating a median. For this loss function the Bayes estimate is just the median of the posterior or predictive distribution for the population median. If one were using squared error loss for estimating the median then the Bayes estimate is just the mean of the predictive distribution for the population median. The admissibility of these estimators under the appropriate loss function follows from a stepwise Bayes argument in the same way as the proof of admissibility for the Basu estimator of the population mean. In Meeden and Vardeman (1991) and Meeden (1993) the following somewhat surprising fact was noted. For many common distributions the mean of the predictive distribution for the population median performed better than the median of the predictive distribution for the population median under both loss functions. Similar results hold for this problem. Hence our estimator will be the mean of the predictive distribution for the population median even though we will follow standard practice and use absolute error as our loss function. We will denote this estimator by *estpp*. This estimator cannot be found explicitly. However we will find it approximately by simulating observations from the posterior or predictive distribution for the population median. Under the Polya sampling scheme for the ratios described above we can simulate a possible realization of the entire population. For this simulated copy we can then find its median. If we repeat this process R times then we have simulated the predictive distribution of the population median under the 'Polya posterior'. When R is large the mean of these R simulated population medians yields, approximately, the estimate *estpp*.

In what follows we will compare the estimator *estpp* to several other estimators. Another estimator we consider is just the sample median of the y_i 's. This ignores the information contained in the auxiliary variable and is used as a bench mark. It will be denoted by *estsm*. Another

estimator is the natural analogue of the ratio estimator of the population mean. This is discussed in Kuk and Mak (1989) and denoted by *estrm*. It is just the ratio of the median of the y values to the median of the x values in the sample multiplied by the median of all the x values in the population. They proposed two other estimators for the median. We will consider just the first one and denote it by *estkm*. This estimator has a plausible intuitive justification and can be found in their paper. Rao, Kovar and Mantel (1990) considered a designed based estimator for the median. We will denote this estimator by *estrk*. Since this estimator can be time consuming to compute we will find it approximately using a method due to Mak and Kuk (1993). Finally we will consider the estimator proposed in Chambers and Dunstan (1986) and denote it by *estcd*. Actually Chambers and Dunstan propose a whole family of estimators and we will only consider one special case which is appropriate when $u_i = \sqrt{x_i}$ in the super population model described at the beginning of this section. We now briefly outline the argument that leads to their estimator of the median. Let F denote the cumulative distribution function associated with the y values of the population. That is F puts mass $1/N$ on each y_i in the entire population. The first step is to get an estimator of $F(t)$ for an arbitrary real number t . If s denotes our sample of size n then given the sample we can write

$$F(t) = N^{-1} \left\{ \sum_{i \in s} \Delta(t - y_i) + \sum_{j \notin s} \Delta(t - y_j) \right\}$$

where $\Delta(z)$ is the step function which is one when $z \geq 0$ and zero elsewhere. Since the first sum in the above expression is known once we have observed the sample, to get an estimate of $F(t)$ it suffices to find an estimate of the second sum. Now under our assumed super population model the population ratios $(y_i - bx_i)/\sqrt{x_i}$ are independent and identically distributed random variables. Since after the sample s is observed a natural estimate of b is $\hat{b} = \sum_{i \in s} y_i / \sum_{i \in s} x_i$ one could act as if the n known ratios $(y_i - \hat{b}x_i)/\sqrt{x_i}$ for $i \in s$ are actual observations from this unknown distribution. Under this assumption, for a fixed t and a fixed unit j not in the sample s an estimate of $\Delta(t - y_j)$ is just the number of the n known ratios incorporating \hat{b} less than or equal to $(t - \hat{b}x_j)/\sqrt{x_j}$ divided by n . Finally if we sum over all the unobserved units j these estimates of $\Delta(t - y_j)$ we then have an estimate for the second sum in the above expression for $F(t)$ which then yields an estimate of $F(t)$. Once we can estimate $F(t)$ for any t by say $\hat{F}(t)$ then the estimate of the population median is $\inf\{t: \hat{F}(t) \geq 0.5\}$.

3. THE POPULATIONS

We will compare these estimators using several different populations. We begin with three actual populations. The

first is a group of 125 American cities. The x variable is their 1960 populations, in millions, while their y variable is the corresponding 1970 populations, again in millions. The second is a group of 304 American counties. The x variable is the number of families in the counties in 1960, while the y variable is the total 1960 population of the county. Both variables are given in thousands. The third population is 331 large corporations. The x variable is their total sales in 1974 and the y variable their total sales in 1975. The sales are given in billions of dollars. We denote these three populations by *ppcities*, *ppcounties* and *ppsalses*. For the three populations the correlations are .947, .998 and .997. These populations were discussed in Royall and Cumberland (1981). Our *ppcounties* is similar to their population Counties60 except we have taken the x variable to be the number of families rather than the number of households.

We have also considered six artificial populations. In each case the auxiliary variable x was chosen first and then the y variable was generated from it. In some cases we followed the super population model described at the beginning of the previous section for some choice of the u_i 's. In some other cases we violated the assumption that conditional on the value x_i the mean of y_i is bx_i . In all cases the errors, the e_i 's, were independent and identically distributed normal random variables with mean zero and variance one.

In the first population, *ppgamma20*, the x_i 's were a random sample from a gamma distribution with shape parameter twenty and scale parameter one. Then given x_i the conditional distribution of y_i was normal with mean $1.2x_i$ and variance x_i , i.e., $u_i = \sqrt{x_i}$.

In the second population, *ppgamma5a*, the x_i 's were ten plus a random sample from a gamma distribution with shape parameter five and scale parameter one. Then given x_i the conditional distribution of y_i was normal with mean $3x_i$ and variance x_i .

In *ppgamma5b* the auxiliary variable was the same as in *ppgamma5a*. Then given x_i the conditional distribution of y_i was normal with mean $3x_i$ and variance x_i^2 .

In *ppstskew* the auxiliary variable was strongly skewed to the right with mean 42.63, median 39.29 and variance 204.59. Then given x_i the conditional distribution of y_i was normal with mean $x_i + 5$ and variance $9x_i$.

In *ppln* the auxiliary variable was a random sample from a log-normal population with mean and standard deviation (of the log) 4.9 and .586 respectively. Then given x_i the conditional distribution of y_i was normal with mean $x_i + 2 \log x_i$ and variance x_i^2 .

In *ppexp* the auxiliary variable was fifty plus a random sample from the standard exponential distribution. Then given x_i the conditional distribution of y_i was normal with mean $80 - x_i$ and variance $(.6 \log x_i)^2$.

All the populations contain 500 units except *ppstskew* which has 1,000. The correlations between the two variables for these last six populations are .76, .87, .41, .61, .58 and $-.28$ respectively.

In most examples where ratio type estimators are used both the y_i 's and x_i 's are usually strictly positive. In population *ppstskew* 13 of the 1,000 units have a y value which is negative. In the original construction of population *ppln* quite a few more of the y values were negative. The population was modified so that all the values are greater than zero.

Note that these populations were constructed under various scenarios for the relationship between the x and y variables. *Ppgamma20* and *ppgamma5a* satisfy the assumptions of the super population model leading to *estcd*, while *ppgamma5b* is consistent with the assumptions underlying *estpp*. In *ppstskew* the conditional variance of y_i given x_i is consistent with *estcd* while for the unmodified *ppln* it was consistent with *estpp*. In both these cases the assumption for the conditional expectation is not satisfied. For the populations *ppcounties*, *ppgamma5a* and *ppln* we have plotted y against x and y/x against x . The results are seen in Figures 1 through 3.

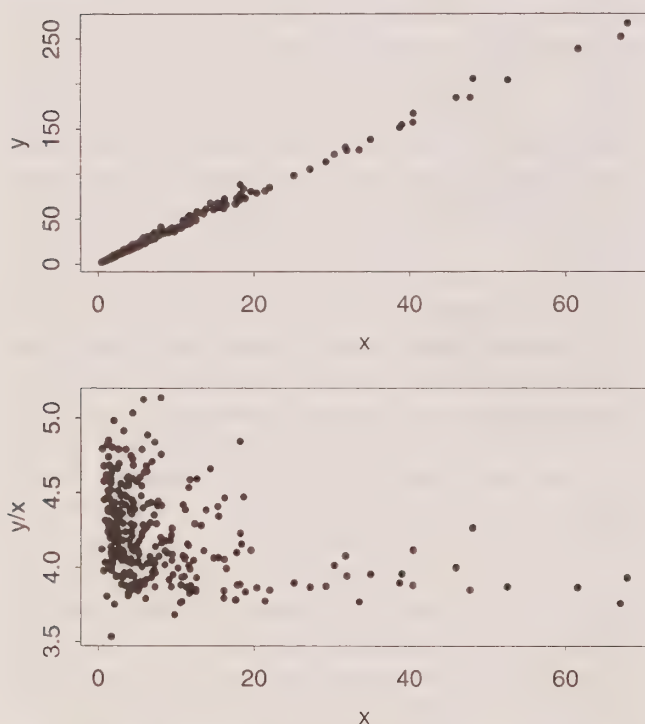


Figure 1. For *ppcounties* the plot of y versus x and y/x versus x where x is the number of families (thousands) living in a county and y is the total population (thousands) of the county for 304 counties.

The estimator *estpp* is based on the assumption that given the sample s our beliefs about the observed ratios, i.e., the ratios y_i/x_i for $i \in s$ and the unobserved ratios, i.e., the ratios y_j/x_j for $j \notin s$ are roughly exchangeable. In particular this means that one's beliefs about a ratio y_j/x_j

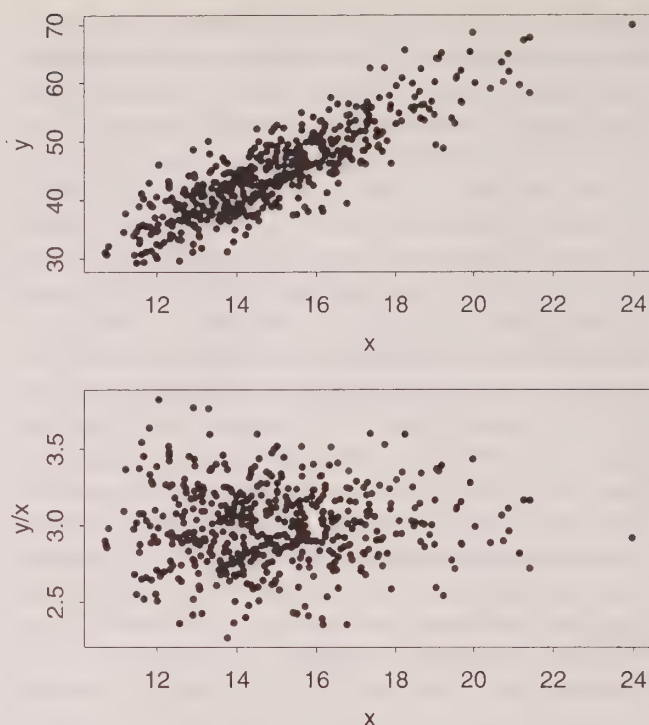


Figure 2. For *ppgamma5a* the plot of y versus x and of y/x versus x .

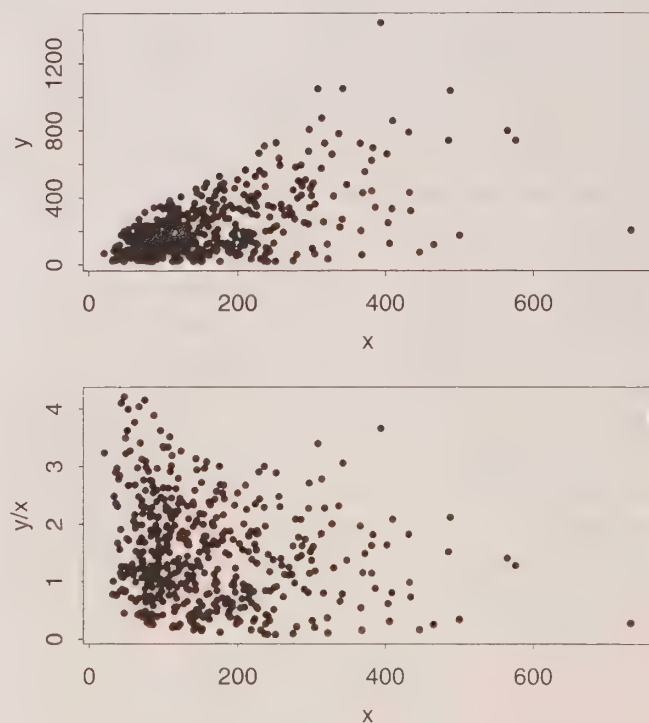


Figure 3. For *ppln* the plot of y versus x and of y/x versus x .

should not depend on the size of x_j . In fact *ppgamma5b* was constructed so that this would indeed be true. On the other hand, under the super population model leading to the estimator *estcd* we would expect the variability of the ratios to get smaller as the size of the x variable increases while the average value of the ratios in any thin vertical strip remains roughly constant as the strip moves to the right. This is seen clearly in the plot of the ratios for population *ppgamma5a*. For the rest of the populations, except for *ppgamma20* the values of the ratios do in fact depend on size of x . This is seen clearly in the plots for *ppcounties* and *ppln*. Hence they should make interesting test cases for the estimator *estpp*. *Ppexp* was included as a test case to see what would happen if the underlying assumptions of *estpp* and *estcd* were strongly violated.

4. SOME SIMULATION RESULTS

To compare the six estimators 500 simple random samples of various sizes were taken from the nine populations. For each sample the values of the six estimators were computed. For the estimator *estpp* this meant finding it approximately by simulating $R = 500$ realizations of the predictive distribution for the population median induced by the ‘Polya posterior’. In each case the average value and average absolute error of the estimator were computed. In Table 1 the average values of all the estimators except *estsm* are given. All the estimators are approximately unbiased except in one case, *estcd* for the population *ppln*. We did not include the results for *estsm* since it is well known that it is unbiased. In Table 2 the average absolute error for all six estimators are given. We see from Table 2 that *estcd* and *estpp* are the clear winners. They both perform better than the other four estimators in every case but one. In *ppexp* they are both beaten by *estsm*, but this is one case where neither would be expected to do well. For the first seven populations their performances are nearly identical while for population *ppln* the estimator *estpp* is preferred and for population *ppstskew* the opposite is true.

In practice one often desires interval estimates as well as point estimates for parameters of interest. Kuk and Mak (1989) and Chambers and Dunstan (1986) each suggested possible methods for finding interval estimates based on their estimator using asymptotic theory. But in each case they did not actually find any interval estimators. Meeden and Vardeman (1991) noted how approximate 95% credible regions based on the ‘Polya posterior’ can be found approximately. If we let $q(.025)$ and $q(.975)$ be the .025 quantile and the .975 quantile of the collection of 500 simulated population medians under the ‘Polya posterior’ then $(q(.025), q(.975))$ is an approximate 95% credible interval. (See Berger 1985 for the definition of such intervals.) Table 3 gives the average length and relative frequency of

coverage for these intervals. We see that for these populations the intervals have reasonable frequentist properties. Perhaps this is not unexpected given the discussion in Meeden and Vardeman (1991). But on the other hand only one of the populations was constructed so that the ratios y_i/x_i are exchangeable. These results suggest that point and interval estimators of the median based on the ‘Polya posterior’ for the ratios are fairly robust against the exchangeability assumption and should work well in a variety of situations. This will be discussed further in section 5.

Table 1
The Average Value of Five Estimators of the Median for 500 Simple Random Samples

Population (median)	Sample Size	Average Value of the Estimator				
		<i>estrm</i>	<i>estkm</i>	<i>estrk</i>	<i>estcd</i>	<i>estpp</i>
<i>ppcities</i> (1.90)	25	.197	.196	.193	.195	.195
<i>ppsals</i> (1.24)	30	1.21	1.25	1.23	1.25	1.24
<i>ppcounties</i> (18.33)	30	18.21	18.60	18.66	18.26	18.39
<i>ppexp</i> (29.02)	30	29.03	29.05	29.00	29.03	29.05
<i>ppgamma5a</i> (43.90)	30	43.82	43.88	43.91	43.99	43.89
	50	43.90	43.91	43.85	44.06	43.90
<i>ppgamma5b</i> (44.17)	30	43.84	43.96	44.19	44.15	43.61
	50	44.28	44.37	44.18	44.18	43.98
<i>ppgamma20</i> (23.15)	30	23.47	23.28	23.14	23.46	23.77
	50	23.34	23.18	23.17	23.43	23.18
<i>ppln</i> (170.25)	30	171.15	169.38	168.12	185.01	170.61
	50	169.15	167.54	167.65	185.03	169.61
<i>ppstskew</i> (46.12)	30	43.66	40.27	45.88	45.50	45.11
	50	44.04	40.70	46.01	45.43	45.37

Table 2
The Average Absolute Error of Six Estimators of the Median for 500 Simple Random Samples

Population	Sample Size	Average Absolute Error of the Estimator					
		<i>estsm</i>	<i>estrm</i>	<i>estkm</i>	<i>estrk</i>	<i>estcd</i>	<i>estpp</i>
<i>ppcities</i>	25	.0326	.0161	.0162	.0155	.0075	.0072
<i>ppsals</i>	30	.1797	.0770	.0797	.0870	.0244	.0245
<i>ppcounties</i>	30	3.12	.586	.964	1.34	.215	.214
<i>ppexp</i>	30	.43	.49	.48	.47	.48	.46
<i>ppgamma5a</i>	30	1.36	.96	1.03	.89	.54	.53
	50	.95	.74	.78	.65	.44	.43
<i>ppgamma5b</i>	30	2.84	2.74	2.71	2.58	2.37	2.38
	50	2.08	2.04	2.01	1.89	1.80	1.85
<i>ppgamma20</i>	30	1.08	1.06	1.05	.88	.67	.64
	50	.94	.77	.78	.73	.51	.49
<i>ppln</i>	30	25.9	25.8	24.2	21.62	21.4	17.0
	50	18.0	20.1	17.9	16.46	17.7	12.7
<i>ppstskew</i>	30	3.86	4.26	6.69	3.21	2.72	3.14
	50	2.92	3.63	5.82	2.55	2.20	2.51

Table 3

The Average Length and Relative Frequency of Coverage for a .95 Credible Interval for the Median Based on the 'Polya Posterior' for 500 Simple Random Samples

Population	Sample Size	Average Length	Frequency of Coverage
<i>ppcities</i>	25	.041	.968
<i>ppsalses</i>	30	.141	.964
<i>ppcounties</i>	30	1.44	.994
<i>ppexp</i>	30	2.26	.944
<i>ppgamma5a</i>	30	2.70	.950
	50	2.15	.956
<i>ppgamma5b</i>	30	11.67	.932
	50	8.86	.942
<i>ppgamma20</i>	30	3.24	.960
	50	2.51	.964
<i>ppln</i>	30	84.8	.934
	50	65.4	.956
<i>ppstskew</i>	30	15.52	.936
	50	12.00	.938

5. DISCUSSION

The motivation for the estimator *estpp* is based on the assumption that the population ratios y_i/x_i 's are exchangeable. This assumption can be described mathematically in two separate but related ways. The first is the super population model given earlier while the second comes from the 'Polya posterior' which is based on a stepwise Bayes argument and gives a non-informative Bayesian interpretation for the estimator. This second approach can be used no matter what parameter is being estimated. When estimating the mean it leads to Basu's estimator which performs very much like the ratio estimator although the ratio estimator usually does a bit better. When estimating the median it leads to the estimator discussed in this note. Here we have argued that the 'Polya posterior' for the ratios leads to good point and interval estimators for the median when an auxiliary variable is present and seems to be reasonably robust against the assumption that the ratios y_i/x_i 's are exchangeable.

Royall and Cumberland (1981) gave an empirical study of the ratio estimator and estimators of its variance. They argued that given a sample an estimate of variance based on the super population model, which leads to the ratio estimator, often made more sense than a design based estimate based on a probability sampling distribution. In Royall and Cumberland (1985), they demonstrated that, conditional on the sample mean of the auxiliary variable, the conditional coverage properties of the usual designed based confidence interval for the population mean were 'hopelessly unreliable'.

We now wish to address the question of the conditional behavior of the intervals for the median based on the Polya posterior which were developed in this note. In the simulation studies given earlier simple random sampling was used for convenience. To get some idea of the conditional behavior of the 'Polya posterior' we considered five of our populations. In each case we ordered the population using the values of the auxiliary variable x . We then took 500 random samples from the first or smallest half of the population, then 500 more random samples from the second or largest half of the population and finally 500 more random samples from the middle third of the population. We then calculated the .95 credible interval for the median based on the 'Polya posterior' which assumes the exchangeability of the ratios y_i/x_i 's. In Table 4 we give the results for the 'Polya posterior' estimators for the median. (We also computed the average value and average absolute error of *estcd* for these examples. We did not include these results since they match closely the results of the 'Polya posterior'.) We see that their conditional behavior, at least in these cases, is very much like their unconditional behavior. In short, interval estimates for the median based on the 'Polya posterior' should have reasonable frequentist properties, no matter how the sample was selected, as long the population approximates our beliefs that the ratios are roughly exchangeable.

Table 4

The Average Value and Absolute Error for the Point Estimator and the Average Length and Relative Frequency of Coverage for a .95 Credible Interval for the Median Based on the 'Polya Posterior' for 500 Simple Random Samples from the whole Population, the 'Smallest' Half, the 'Largest' Half and the 'Middle' Third

Population	Sample Size	Where Taken	Average Value	Average Error	Average Length	Frequency of Coverage
<i>ppcities</i>	25	whole	.195	.0072	.041	.968
		smallest ½	.192	.0047	.033	.994
		largest ½	.196	.0078	.048	.988
		middle ⅓	.201	.0114	.055	.922
<i>ppcounties</i>	30	whole	19.4	.220	1.46	.990
		smallest ½	18.6	.305	1.34	.942
		largest ½	18.1	.283	1.59	.954
		middle ⅓	18.5	.252	1.35	.964
<i>ppsalses</i>	30	whole	1.24	.0072	.141	.964
		smallest ½	1.24	.027	.153	.966
		largest ½	1.23	.020	.125	.982
		middle ⅓	1.23	.027	.139	.944
<i>ppgamma5a</i>	30	whole	43.9	.53	2.70	.950
		smallest ½	43.8	.55	2.82	.948
		largest ½	44.0	.53	2.55	.940
		middle ⅓	43.9	.47	2.63	.974
<i>ppgamma5b</i>	30	whole	43.6	2.38	11.7	.932
		smallest ½	42.2	2.69	11.6	.890
		largest ½	45.1	2.25	11.2	.950
		middle ⅓	45.2	2.27	11.3	.936

As can be seen by looking at the plots of y_i/x_i versus x_i and our simulation results it does not seem to matter much if the variability in the ratios y_i/x_i 's decreases as x_i increases. What is crucial however is that the average value of the ratios in the narrow strip above a small interval of possible x values remains fairly constant as we move the small interval to the right. In Figure 2, the plot of the ratios for *ppgamma5a* is an example of such a plot. In fact this is how the population was constructed, since it satisfies the assumptions underlying *estcd*. In Figures 1 and 3 we see for *ppcounties* and *ppln* that the average value of the ratios in a narrow strip tends to decrease as we move to the right and helps to explain the relatively poorer performance of the 'Polya posterior' estimators in these cases. Overall however, the performance of procedures based on the 'Polya posterior' seem to be reasonably robust against the exchangeability assumption.

As another alternative we could consider a more balanced sampling plan which is based on stratifying the population on the auxiliary variable. For example consider again population *ppgamma5b* where it is ordered on the basis of its x_i values. We constructed ten strata where the first stratum consisted of the units with the fifty smallest x_i values, the second stratum of the units with the next fifty smallest x_i values and so on. We then took 500 stratified random samples of size fifty where five units were chosen at random from each stratum. For these samples the average value of *estpp* was 43.94 and its average absolute error was 1.81. The average length of its corresponding interval estimator was 8.95 with .938 relative frequency of covering the true value. Note that these figures are very similar to those given Tables 1 and 2 when simple random sampling was used.

ACKNOWLEDGEMENTS

Research supported in part by NSF grant SES 9201718.

REFERENCES

- BASU, D. (1971). An essay on the logical foundations of survey sampling, part one. In *Foundations of Statistical Inference*. Toronto: Holt, Reinhart and Winston, 203-242.
- BERGER, J.O. (1985). *Statistical Decision and Bayesian Analysis*. New York: Springer-Verlag.
- BINDER, D.A. (1982). Non-parametric Bayesian models for samples from finite populations. *Journal of the Royal Statistical Society, Series B*, 44, 388-393.
- CHAMBERS, R.L., and DUNSTAN, R. (1986). Estimating distribution functions from survey data. *Biometrika*, 73, 597-604.
- FERGUSON, T.S. (1973). A Bayesian analysis of some non-parametric problems. *Annals of Statistics*, 1, 209-230.
- KUK, A.Y.C., and MAK, T.K. (1989). Median estimation in the presence of auxiliary information. *Journal of the Royal Statistical Society, Series B*, 51, 261-269.
- MAK, T.K., and KUK, A. (1993). A new method for estimating finite-population quantiles using auxiliary information. *The Canadian Journal of Statistics*, 21, 29-38.
- MEEDEN, G., and GHOSH, M. (1983). Choosing between experiments: applications to finite population sampling. *Annals of Statistics*, 11, 296-305.
- MEEDEN, G., and VARDEMAN, S. (1991). A noninformative Bayesian approach to interval estimation in finite population sampling. *Journal of the American Statistical Association*, 86, 972-980.
- MEEDEN, G. (1993). Noninformative nonparametric Bayesian estimation of quantiles. *Statistics and Probability Letters*, 16, 103-109.
- RAO, J.N.K., KOVAR, J.G., and MANTEL, H.J. (1990). On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika*, 77, 365-375.
- ROYALL, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57, 377-387.
- ROYALL, R.M., and CUMBERLAND, W.D. (1981). An empirical study of the ratio estimator and estimators of its variance. *Journal of the American Statistical Association*, 76, 66-88.
- ROYALL, R.M., and CUMBERLAND, W.D. (1985). Conditional coverage properties of finite population confidence intervals. *Journal of the American Statistical Association*, 80, 355-359.
- RUBIN, D.B. (1981). The Bayesian bootstrap. *Annals of Statistics*, 9, 130-134.

Outlier Robust Horvitz-Thompson Estimators

BEAT HULLIGER¹

ABSTRACT

The Horvitz-Thompson estimator (HT-estimator) is not robust against outliers. Outliers in the population may increase its variance though it remains unbiased. The HT-estimator is expressed as a least squares functional to robustify it through M-estimators. An approximate variance of the robustified HT-estimator is derived using a kind of influence function for sampling and an estimator of this variance is developed. An adaptive method to choose an M-estimator leads to minimum estimated risk estimators. These estimators and robustified HT-estimators are often more efficient than the HT-estimator when outliers occur.

KEY WORDS: Outlier; M-estimator; Adaption; Population mean; Sampling; Sensitivity curve.

1. INTRODUCTION

The mean of a variable over a finite population is an important indicator. Examples are the mean salary of employees in a branch of the economy or the mean yield of corn of the farms in a region. Due to its connection to the sum the mean cannot be easily replaced by other indicators. But the population mean is a sensitive characteristic because a single large observation may determine its value. The Horvitz-Thompson estimator (HT-estimator) is a natural estimator of the population mean if the sample design has unequal inclusion probabilities and is without replacement. It is the sample mean in simple random sampling. It is always unbiased whatever the population distribution of the investigated variable is. But the HT-estimator is not robust against outliers because it is linear in the observed values like its estimand, the population mean. Large observations together with small inclusion probabilities have a particularly large influence on the HT-estimator.

Suppose there is an outlier in a sample. The outlier may be a correct observation from the target population. Discarding such a correct outlier makes the HT-estimator biased. But keeping it with full weight makes the HT-estimator highly variable because typically the outlier would show up only in a few of the possible samples. Thus there is a tradeoff between bias and variance in this case, which in particular includes asymmetric distributions with one heavy tail.

The outlier may also be an incorrect observation, *e.g.*, due to a measurement or coding error or stemming from an element outside the target population. In that case keeping the outlier with full weight may entail a large bias of the HT-estimator in addition to high variability. Thus discarding incorrect outliers reduces both bias and variance.

Since it is often difficult to detect outliers and to decide whether it is correct or not one would like to have estimators that perform well in terms of bias and variance

irrespective of the nature and the detection of possible outliers. HT-estimators which are robustified through M-estimators are promising candidates for this difficult task.

In the survey sampling literature the problem of outliers or aberrant values is often treated under the heading “skew populations”. Kish (1965, sec. 11.4 B) describes the problem in economical surveys and surveys of individuals. He proposes the formation of separate strata for outliers if possible, truncation, transformation or modelling. The idea of forming a separate class for large units and combining the class means is investigated for example in (Glasser 1962) and (Hidioglou and Srinath 1981).

The truncation idea is made more precise by the winsorized mean proposed by Searls (1966). Fuller (1991) proposed a preliminary-test-estimator which reduces the impact of the largest data values only when a test for extreme values is significant. Rivest (1993) studied the behavior of various winsorization schemes under simple random sampling. Shoemaker and Rosenberger (1983) derive exact formulae for the expected value and variance of the median and trimmed mean under simple random sampling without replacement. Oehlert (1985) proposes the random average mode estimator to estimate the mean of finite populations in an outlier robust way. Smith (1987) emphasises that it is as important to detect and treat influential observations if the inference is based on the randomisation provided by the sample design as if the observations are considered realisations of random variables. He proposes an influence measure for linear estimators based on case deletion, which involves both the variable of interest and its weight.

The prediction approach in sampling theory uses stochastic models for the population to predict the total of the present realisation. Linear models and (nonrobust) linear estimators are used. Aspects of the sensitivity and robustification against model misspecification are reviewed

¹ Beat Hulliger, Swiss Federal Statistical Office, Schwarztorstrasse 96, CH-3003 Bern, Switzerland.

in (Iachan 1984). Chambers (1986) develops an outlier-robustification of the prediction approach using M-estimators. He distinguishes representative and nonrepresentative outliers in a sample. Representative outliers must be included with full weight in an unbiased estimate of the population mean while nonrepresentative outliers should be downweighted or discarded.

Little and Smith (1987) treat outliers and missing data in certain positive multivariate continuous data by a robustified EM-algorithm. Gwet and Rivest (1992) investigate resistant ratio estimators under simple random sampling without replacement.

M-estimators form a class of flexible and simple robust estimators. An M-estimator T of location is defined implicitly by the estimating equation

$$\sum_{i=1}^n \psi(X_i - T) = 0$$

for a predetermined function ψ , e.g., $\psi_{\text{Hub}}(x, k) = \max(-k, \min(k, x))$, where k is a tuning constant. An M-estimator may be written as a functional of the empirical distribution function. The influence function of an estimator is a functional derivative of the estimator (Hampel 1974). It describes the reaction of the estimator to a small contamination in the data. An M-estimator with bounded ψ -function usually has a bounded influence function such that outliers cannot disturb the estimator too much. For the estimation of the mean of asymmetric finite populations M-estimators must be adapted.

In this article we develop design-based M-estimators for samples with unequal inclusion probabilities. The simple linear model which implicitly is the basis of the Horvitz-Thompson strategy is made explicit and the HT-estimator is expressed as a functional of an empirical distribution function which accounts for the complex sample design. This establishes the link to classical robust statistics and allows a straightforward robustification of the HT-estimator (Section 2). We define an influence function for sampling which clarifies the outlier-sensitivity of the HT-estimator and leads to an approximation of the sampling variance of the robustified HT-estimator. An estimator of this variance is presented. In Section (3) we briefly comment on stratification, domains, robust designs and one-step estimators. In Section (4) an adaptive robustification of the HT-estimator is developed. The method chooses from a class of robustified HT-estimators the one which minimizes an estimate of the mean squared error. The resulting estimator is called minimum estimated risk estimator (MER-estimator). A Monte-Carlo simulation is presented in Section 5. Robustified HT-estimators and MER-estimators outperform the HT-estimator in many outlier situations. The premium to pay is a moderate loss of efficiency in situations where the HT-estimator is optimal.

2. ROBUSTIFICATION OF HORVITZ-THOMPSON ESTIMATORS

2.1 The HT-Estimator as a Least Squares Functional

A finite population $U = \{1, \dots, N\}$ of $0 < N < \infty$ distinct elements is sampled. We are interested in a variable y which takes the values y_i for $i \in U$. The sample design $p(S)$ on the space of samples S of fixed size n has inclusion probabilities $\pi_i = P[i \in S] = \sum_{S \ni i} p(S)$. These π_i are proportional to some known positive auxiliary variable x_i ($i \in U$). Such sample designs are called IPPS designs (inclusion probability proportional to size) because often x_i is some size measure. Denote by π_{ij} the joint inclusion probability $P[i \in S, j \in S]$ ($i, j \in U$). The vector of all y -values is denoted $y_U = (y_1, \dots, y_N)^T$ and x_U is defined in an analogous way. The vector of the y -values of a sample S is denoted $y_S = (y_{i_1}, \dots, y_{i_n})^T$ ($i_k \in S$). The goal is to estimate the population mean of the variable y : $\bar{y}_U = \sum_{i \in U} y_i / N$.

The HT-estimator for \bar{y}_U is $T_{HT} = \sum_{i \in S} y_i / (N\pi_i)$. The variance of T_{HT} is estimated by the well known estimator

$$v_{HT}(T_{HT}) =$$

$$\frac{1}{N^2} \left[\sum_{i \in S} (1 - \pi_i) \frac{y_i^2}{\pi_i^2} + \sum_{i \neq j \in S} (1 - \pi_i \pi_j / \pi_{ij}) \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} \right], \quad (1)$$

which is due to Horvitz and Thompson or by the variance estimator due to Yates, Grundy and Sen (see Cochran 1977, p. 261).

The rationale behind the HT-estimator given in the survey sampling literature is that it has sampling variance zero if the inclusion probabilities π_i are exactly proportional to y_i . Then $T_{HT}(y_S) = \bar{y}_U$ for every sample S . The HT-estimator is bias-robust but not variance-robust with respect to deviations from proportionality between y_i and π_i (cf. Rao 1966).

How can the HT-estimator be formulated in a way which allows the derivation of an influence function analogue and a variance estimator? The key idea is to express the HT-estimator as a least squares (LS) functional of an estimate of the population distribution function in such a way that the design is incorporated in the estimator of the population distribution function while the proportionality of y_i and x_i is taken up by the LS-functional.

The joint population distribution function of two variables (x_i, y_i) is defined as $F_U(r, t) = \sum_{i \in U} \mathbf{1}\{x_i \leq r\} \mathbf{1}\{y_i \leq t\} / N$, where $\mathbf{1}\{y_i \leq t\} = 1$ if $y_i \leq t$ and 0 elsewhere. There are various possibilities to estimate F_U but the easiest and most generally applicable estimator is the sample distribution function

$$F_S(r, t) = \sum_{i \in S} \frac{1}{\pi_i} \mathbf{1}\{x_i \leq r\} \mathbf{1}\{y_i \leq t\} \Big/ \sum_{i \in S} \frac{1}{\pi_i}. \quad (2)$$

The estimator F_S is a distribution function itself.

To derive a LS-functional the following superpopulation model for the proportionality between y_i and x_i is used: We assume that y_U is a vector of realisations of independent random variables Y_i with expectation βx_i and variance $\sigma^2 x_i$.

Definition 1. The LS-estimator $\beta_{LS}(F_S)$ of β in the above model with respect to the sampling distribution function F_S of (x_i, y_i) ($i \in S$) minimizes $\int (y - \beta x)^2 / x dF_S(x, y)$ or equivalently solves

$$\sum_{i \in S} \frac{1}{\pi_i} \left(\frac{y_i - \beta x_i}{\sqrt{x_i}} \right) \frac{x_i}{\sqrt{x_i}} = 0. \quad (3)$$

The following statement is well known and its proof is easy. If S is a sample drawn according to an IPPS sample design with inclusion probabilities $\pi_i = n x_i / \sum_{i \in U} x_i$ ($i \in U$) then the HT-estimator is $T_{HT} = \bar{x}_U \beta_{LS}(F_S)$, where $\beta_{LS}(F_S)$, the LS-estimator defined by (3); is given by

$$\beta_{LS}(F_S) = \frac{\sum_{i \in S} y_i / \pi_i}{\sum_{i \in S} x_i / \pi_i}.$$

Note that the expression $T_{HT} = \bar{x}_U \beta_{LS}(F_S) = \bar{x}_U (\sum_{i \in S} y_i / \pi_i) / (\sum_{i \in S} x_i / \pi_i)$ does not depend on the superpopulation model. However the superpopulation model clarifies the role of the HT-estimator: The slope $\beta_{LS}(F_S)$ involved in the HT-estimator is a weighted least squares estimator that incorporates the information in the design through F_S as well as the information in the auxiliary variable through the regression.

2.2 The Robustified HT-Estimator

After the separation of design and auxiliary information and its expression as a LS-functional the robustification of the HT-estimator is analogous to the robustification of LS-estimators in linear models for infinite populations through M-estimators (cf. Hampel *et al.* 1986, Chapter 6): The estimating equation (3) now involves some function η which depends on the standardized residuals $(y_i - \beta x_i) / x_i^{1/2}$ and on x_i . For ease of notation denote by a prime the division by $x_i^{1/2}$ and let $r'(\beta) = (y - \beta x) / x^{1/2}$.

Definition 2. Let $\beta(F_S, \eta)$ be a solution of the equation

$$\sum_{i \in S} \frac{1}{\pi_i} \eta(x'_i, r'_i(\beta)) x'_i = 0. \quad (4)$$

The robustified HT-estimator (RHT-estimator) is

$$T_{RHT}(F_S) := \bar{x}_U \beta(F_S, \eta).$$

$\beta(F_S, \eta)$ is called the slope of the RHT-estimator.

In general useful choices of η are of the form $\eta(x, r) = w(x) \psi(r \cdot u(x))$, where $w(x)$ and $u(x)$ are two weighting functions and ψ is a defining function for a location M-estimator (cf. Hampel *et al.* 1986, p. 315). In the following we use the so-called Mallows form, which sets $u(x) \equiv 1$. Mallows-type regression downweights outlying x -values and outlying residuals independently. A well-known example, which also sets $w(x) \equiv 1$, is the Huber-function $\eta(x, r) = \psi_{\text{Hub}}(r, k) = \max(-k, \min(k, r))$ for some constant k . The RHT-estimator with defining function $\eta(x, r) \equiv r \forall x$ is the HT-estimator. Thus by adjusting the tuning constant k in the Huber-function a smooth transition of estimators from the HT-estimator to more and more robust estimators is possible.

Scale estimates are needed in $w(x)$ and $\psi(r)$ to make $\beta(F_S, \eta)$ scale equivariant. While for the weighting function $w(x'_i)$ preliminary scale estimators are available, e.g., the median of the x'_i , the scale of the residuals must be estimated simultaneously with the slope β . The median of the absolute residuals may be used. In the following theoretical development (Sections 2.3 to 4) scale is assumed known to simplify the treatment.

The RHT-estimator is a nonparametric estimator. The model $Ey = \beta x$ is merely used to motivate the expression of the HT-estimator as a least squares functional. Neither the HT-estimator nor the RHT-estimator need this model or symmetry of errors with variance proportional to x in order to be applied.

Other formulations of the HT-estimator as least squares functionals may be appropriate in certain conditions. Suppose that in spite of the IPPS-design y_i is not correlated with π_i . Then one would probably choose the unweighted sample mean $\bar{y}_S = \sum_{i \in S} y_i / n$ as an estimator of the population mean (cf. Rao 1966). A robustification of \bar{y}_S could be a solution $\hat{\mu}$ of $\sum_{i \in S} \psi(y_i - \mu) = 0$. This is a location M-estimator. If the HT-estimator is in fact appropriate due to the correlation between y_i and π_i then this robustification is not efficient.

A third robustification would assume y_i proportional to x_i but with variance proportional to the square of x_i . This is in fact the situation where the HT-estimator is optimal. The corresponding robustification would be a solution $\hat{\beta}$ of $\sum_{i \in S} \eta(x_i, y_i / x_i - \beta) = 0$. Obviously this robustification does not account for the IPPS-sample design. If the design is put back into the estimating equation by solving $\sum_{i \in S} \eta(x_i, y_i / x_i - \beta) / \pi_i = 0$ then we do not get back the HT-estimator when $\eta(x, r) \equiv r$.

One may argue that in fact the HT-estimator is never used in its pure form for estimating population means. The usual estimator is $(\sum_{i \in S} y_i / \pi_i) / (\sum_{i \in S} 1 / \pi_i)$, sometimes called the Hájek-estimator. The estimating equation of the Hájek-estimator, $\sum_{i \in S} (y_i - T) / \pi_i = 0$, makes obvious that the Hájek estimator is not robust against outliers in y .

But the residual $y_i - T$ does not involve the auxiliary variable x_i . Therefore the Hájek-estimator does not suffer from a possible combined effect of large y_i together with small x_i , which may be a leverage point for the regression model underlying the HT-estimator.

2.3 A Sampling Sensitivity Curve

The derivation of an approximate sampling variance of the RHT-estimator (see Section 2.4) uses a finite population analogue to the influence function for infinite populations (Hampel 1974). For finite population sampling with design based inference it is appropriate to develop a sensitivity curve (SC) (cf. Hampel *et al.* 1986, p. 93) for $\beta(F, \eta)$ at the population distribution function F_U . In other words, the slope of the RHT-functional is linearized around F_U . Denote by $U+$ the population U augmented by a unit with characteristic (x, y) . Denote by $\lambda(\beta, F_U)$ the function $\sum_{i \in U} \eta(x'_i, r'_i(\beta)) x'_i / N$, such that the defining equation for $\beta(F_U, \eta)$, the M-estimator at the population distribution function, is $\lambda(\beta, F_U) = 0$. Clearly

$$(N + 1) [\lambda(\beta(F_{U+}, \eta), F_{U+}) - \lambda(\beta(F_U, \eta), F_U)] = 0.$$

Using a linear approximation to $\eta(x, \cdot)$ and neglecting terms in $1/N$ the sensitivity curve of $\beta(F_U, \eta)$ can be isolated from this equation:

$$(N + 1) (\beta(F_{U+}, \eta) - \beta(F_U, \eta)) \approx \frac{\eta(x', r') x'}{\sum_{i \in U} \eta_2(x'_i, r'_i) x'^2_i / N} =: SC(x, y, F_U, \eta), \quad (5)$$

where $\eta_2(x, r) = \partial \eta(x, r) / \partial r$ and both r' and r'_i are evaluated at $\beta(F_U, \eta)$. This SC may be extended to the case of a p -dimensional explanatory variable (cf. Hampel *et al.* 1986, p. 316 and Hulliger 1991, p. 183).

Since units usually are not independently included into an IPPS sample, the reaction of the RHT-slope to a particular observation must be investigated by conditioning on a particular sample. The deviation of the estimator $\beta(F_S, \eta)$ at a particular sample S from $\beta(F_U, \eta)$ may be approximated by integrating the SC of $\beta(F, \eta)$ with respect to the sampling distribution function F_S (cf. Hampel *et al.* 1986, p. 85):

$$\beta(F_S, \eta) - \beta(F_U, \eta) \approx \int SC(x, y, F_U, \eta) dF_S. \quad (6)$$

The influence of unit i in sample S may then be defined as the contribution of the unit i to the deviation due to the sample S , i.e.,

$$SC((x_i, \pi_i, y_i) \mid S, F_U, \eta) = \frac{\eta(x'_i, r'_i) x'_i / \pi_i}{(\sum_{j \in S} 1 / \pi_j) \sum_{j \in U} \eta_2(x'_j, r'_j) x'^2_j / N}. \quad (7)$$

The SC may be studied theoretically to discuss the properties of the RHT-estimator and to choose a good η -function. And it may be estimated by replacing the standardization factor $N / (\sum_{j \in U} \eta_2(x'_j, r'_j) x'^2_j)$ by an appropriate estimator. The estimated SC may be used as a tool for outlier detection.

The influence of unit i in sample S on the HT-estimator is

$$\bar{x}_U SC((x_i, \pi_i, y_i) \mid S, F_U, \eta \equiv r) = (y_i - \beta_{LS}(F_U) x_i) \left/ \left(1 + \pi_i \sum_{j \in S \setminus i} 1 / \pi_j \right) \right.$$

This SC is unbounded in y_i such that the HT-estimator is not robust against outlying y_i . The y_i influences the HT-estimator through the residual $y_i - \beta_{LS}(F_U) x_i$. This makes clear why a large y_i combined with a small x_i (or small π_i) has a large influence. If π_i is directly proportional to x_i , as the IPPS design in principle requires, then the SC of the HT-estimator is bounded in x_i . In other words the HT-estimator is robust against outlying x_i . However the bound may be quantitatively too high to be efficient and further downweighting of outlying x_i may be necessary.

2.4 Approximate Expectation and Variance

Along the lines of the proof of proposition 2.1 in Gwet and Rivest (1992) it can be shown that $\beta(F_S, \eta)$ is consistent for $\beta(F_U, \eta)$ in the sense that for a growing and nested sequence of populations and IPPS samples $\lim_{N, n \rightarrow \infty} P[|\beta(F_S, \eta) - \beta(F_U, \eta)| < \epsilon] = 1 \forall \epsilon > 0$.

Due to the consistency of $\beta(F_S, \eta)$ the sampling expectation $E_S \beta(F_S, \eta)$ is approximately $\beta(F_U, \eta)$. Of course $\bar{x}_U \beta(F_U, \eta)$ may be different from the population mean and then $\bar{x}_U \beta(F_S, \eta)$ has a bias as an estimator of \bar{y}_U . In particular if the population distribution is not symmetric then $\bar{x}_U \beta(F_S, \eta)$ is in general a biased estimator for \bar{y}_U but nevertheless consistent for $\bar{x}_U \beta(F_U, \eta)$. The important question then is how large is the bias of $\bar{x}_U \beta(F_S, \eta)$, in particular when compared with the variance.

The SC (5) may be used to derive a variance approximation. The derivation is analogous to the case of independent identically distributed random variables with the influence function replaced by the sampling SC. Taking the expectation of the square of (6) one gets after some approximations

$$\begin{aligned} \text{Var}_S \beta(F_S, \eta) &\approx E_S [(\beta(F_S, \eta) - \beta(F_U, \eta))^2] \\ &\approx \frac{\text{Var}_S (\sum_S \eta(x'_i, r'_i) x'_i / \pi_i)}{(\sum_{i \in U} \eta_2(x'_i, r'_i) x'^2_i)^2} \end{aligned}$$

$$\approx \frac{\sum_{i \in U} \left(\frac{1}{\pi_i} - 1 \right) \eta(x'_i, r'_i)^2 x_i'^2 + \sum_{i \neq j \in U} \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) \eta(x'_i, r'_i) x'_i \eta(x'_j, r'_j) x'_j}{\sum_{i \in U} \eta_2(x'_i, r'_i)^2 x_i'^4 + \sum_{i \neq j \in U} \eta_2(x'_i, r'_i) x_i'^2 \eta_2(x'_j, r'_j) x_j'^2}, \quad (8)$$

where r'_i is evaluated at $\beta(F_U, \eta)$. Denote this approximate variance by V_r . An important difference to the case of the asymptotic variance of an M-estimator with independent identically distributed random variables is that the cross-product terms in the numerator of V_r do not vanish. If $\eta(x, r) \equiv r$ then V_r yields the correct variance of the HT-estimator.

2.5 Estimation of the Variance

The numerator of V_r is the variance of $\sum_{i \in S} \eta(x'_i, r'_i) (\beta(F_U, \eta) x'_i / \pi_i)$ which is a HT-estimator apart from the unknown r'_i ($\beta(F_U, \eta)$). Therefore the variance estimator (1) for the HT-estimator may be used. After replacing $\beta(F_U, \eta)$ by the estimator $\beta(F_S, \eta)$, the estimator of the variance of the RHT-estimator becomes

$$v_{rHT} = -\bar{x}_U^2 \frac{\sum_{i \in S} \frac{1}{\pi_i} \eta(x'_i, r'_i)^2 x_i'^2 + \sum_{i \neq j \in S} \frac{1}{\pi_{ij}} \eta(x'_i, r'_i) x'_i \eta(x'_j, r'_j) x'_j}{\sum_{i \in S} \frac{1}{\pi_i} \eta_2(x'_i, r'_i)^2 x_i'^4 + \sum_{i \neq j \in S} \frac{1}{\pi_{ij}} \eta^2(x'_i, r'_i) x_i'^2 \eta_2(x'_j, r'_j) x_j'^2}. \quad (9)$$

The minus sign in (9) is in order. The (negative) cross-product terms in the numerator usually dominate. Nevertheless v_{rHT} may become negative as can the HT-variance estimator (1) itself (cf. Cochran 1977, p. 261). The variance estimator v_{rHT} does not yield the variance estimator (1) if $\eta(x, r) \equiv r$. Of course the Yates-Grundy-Sen estimator may be used to estimate the numerator of V_r . A third variance estimator may be derived by writing the RHT-estimator as a weighted least squares estimator whose weights depend on the estimate (cf. Hulliger 1991, p. 166). Since the MER-estimators (cf. Section 4) performed slightly better with v_{rHT} than with the other variance estimators the simulations of Section 5 were done with v_{rHT} .

3. EXTENSIONS

3.1 Stratification and Domains

The stratified mean under stratified random sampling is a HT-estimator. The stratified mean may be written as the mean of predicted values under a one-way analysis of variance model. The corresponding robustification is straightforward. It amounts to the separate robustification of the stratum means (Hulliger 1991). However, if the stratum sample size is 1 or 2 no outlier can be down-weighted without the help of further assumptions. Furthermore the biases of the robustified stratum means may add up to a large overall bias (cf. Rivest 1993, Section 4).

Therefore different robustifications may be appropriate for estimating stratum means and overall means.

This is a general problem for robust estimation in subpopulations (domains) since the definition of an outlier depends on the reference population. An observation may be an outlier in a particular subpopulation but may be harmless in another one. Thus a robust estimator may be suited for one subpopulation but perform poorly in another subpopulation. Often no robustification is needed or wanted for overall means but subpopulation means need to be robustified because of outliers that turn up. Luckily the sample size is often considerably smaller in a subpopulation than in the whole population and then the bias component of the MSE of a robust estimator is often smaller than the variance component. Thus robust estimators may be more efficient than the HT-estimator when used in domain estimation.

3.2 Hansen-Hurwitz Strategy

When sampling is done with replacement and with unequal drawing probabilities the Hansen-Hurwitz estimator is used instead of the HT-estimator. The Hansen-Hurwitz estimator may be robustified analogously to the HT-estimator (see Hulliger 1991, section 4.4) since the underlying model is the same. The variance approximation for the robustified HH-estimator is simpler than for the RHT-estimator because the crossproduct terms vanish due to the drawing with replacement of the Hansen-Hurwitz design.

3.3 Robustified IPPS Design

The ratios y_i / π_i in the HT-estimator act like the summands of an arithmetic mean. Small π_i together with large y_i inflate the HT-estimator. To robustify the design against very large and very small inclusion probabilities we may put $\tilde{\pi}_i = n \tilde{x}_i / \sum_U \tilde{x}_i$, where $\tilde{x}_i = \tilde{x}_U + \psi_{\text{Hub}}(x_i - \tilde{x}_U, k)$. Thus the auxiliary variable x_i is "Huberised" from its mean to prevent too high and too low values. Now an IPPS sample is drawn with inclusion probabilities $\tilde{\pi}_i$. The HT-estimator is still $T_{HT} = (1/N) \sum_S y_i / \tilde{\pi}_i$ and it is still unbiased. Of course it is not robust against outliers in y and it may lose efficiency if the expectation of the y_i is not proportional to $\tilde{\pi}_i$. The weighted LS-estimator under the superpopulation model for the HT-estimator (see Section 2.1) with inclusion probabilities $\tilde{\pi}_i$ and unmodified auxiliary variable x_i is

$$\beta_{LS}(F_S) = \frac{\sum_S y_i / \bar{\pi}_i}{\sum_S x_i / \bar{\pi}_i}, \quad (10)$$

with corresponding estimator for the population mean $\bar{x}_U \beta_{LS}(F_S)$. This β_{LS} may be robustified against outliers in y_i like the HT-estimator. Ratio estimators in IPPS samples are of the same form with the original π_i instead of $\bar{\pi}_i$. Thus ratio estimators may be robustified analogously to HT-estimators, too (cf. Gwet and Rivest 1992).

3.4 One-step Estimators

It is not advisable to express robust estimators as weighted means with fixed weights attached to the observations because the notion and the effect of an outlier depend on the particular domain and variable to be analysed. However, so-called one-step estimators, which are expressed as weighted means, reduce the computational complexity of robust estimators. The one-step RHT-estimator is

$$\bar{x}_U \frac{\sum_{i \in S} w_i y'_i x'_i / \pi_i}{\sum_{i \in S} w_i x_i'^2 / \pi_i}, \quad (11)$$

with weights $w_i = \eta(x'_i, y'_i - \beta_{LS} x'_i) / (y'_i - \beta_{LS} x'_i)$. In fact this is the result of the first step of the iteratively reweighted least squares algorithm, which is often used to calculate M-estimators. The one-step RHT-estimator inherits much of the good properties of the fully iterated RHT-estimator and is simpler to implement and faster to compute.

4. MINIMUM ESTIMATED RISK ESTIMATORS

The RHT-estimator is in general biased. A convenient performance criterium is the sampling mean squared error (MSE) $E_S[(\bar{x}_U \beta(F_S, \eta) - \bar{y}_U)^2]$. For small to moderate samples the gains of RHT-estimators over the HT-estimator are not very sensitive to the particular robustification chosen if there are outliers in the sample (cf. Hulliger 1991, Chapter 3). But with well-behaved data or for moderate to large samples the losses in MSE of certain RHT-estimators may be considerable. The question arises how to choose a good RHT-estimator. Minimum estimated risk estimators (MER-estimators), which adapt the tuning constant of a RHT-estimator to the sample, are a possibility. MER-estimators for the expectation of a univariate random variable are investigated in Hulliger (1991, Chapter 2). The idea is to take a simple M-estimator like a Huber M-estimator, to estimate its MSE for a set of tuning constants k , and to choose the tuning constant with least estimated MSE.

Huber's (1964, p. 97) proposal 3 and Jaeckels (1971) adaptive trimmed mean aim at symmetric random variables and therefore use a variance estimate instead of an estimate

of the MSE. MER-estimators are similar but their aim is to estimate the mean of asymmetric distributions.

Here we introduce MER-estimators for IPPS designs. Consider a parametric set of functions $\{\eta_k(x, r) : k \in K\}$, where $K \subset \mathbf{R}_+^p$ is the set of parameters. Usually $p = 1$ or 2 to make minimization feasible and to keep the efficiency loss due to the estimation of the nuisance parameter k low. We do not call k a parameter but a tuning constant to avoid any confusion with the concept of parameters in probability distributions. A suitable set of η -functions induces a set $\mathcal{B} := \{\beta(F_S, \eta_k) : k \in K\}$, where $\beta(F_S, \eta_k)$ is the slope of an RHT-estimator. To ensure consistency of the MER-estimator let $\lim_{k \rightarrow \infty} \eta_k(x, r) = r \forall (x, r)$ such that the HT-estimator is an element of \mathcal{B} . The MSE of $\beta(F_S, \eta_k)$ may be estimated by

$$r(F_S, k) = \max(v_r(F_S, k), 0) + (\beta(F_S, k) - \beta_{LS}(F_S))^2, \quad (12)$$

where $v_r(F_S, k)$ is the variance estimator (9) or some other estimator of the variance of $\beta(F_S, \eta_k)$. We use $\max(v_r, 0)$ in $r(F_S, k)$ because the variance estimator (9) may become negative. Typically the function $r(F_S, k)$ with $k \in \mathbf{R}_+$ has a maximum at or close to $k = 0$ which stems from a large bias. Then it drops to a minimum where bias and variance are both small. For large tuning constants $r(F_S, k)$ approaches the variance of the HT-estimator, usually from below.

Definition 3. Suppose $r(F_S, \cdot)$ has a global minimum at $k_m(F_S) \in K$. Then the MER-estimator of the population mean is $M(F_S) = \bar{x}_U \beta(F_S, \eta_{k_m})$.

MER-estimators with suitable defining functions are scale equivariant and do not need a scale estimator. MER-estimators are in general consistent estimators of the population mean. A proof of the strong consistency of MER-estimators of the expectation of a random variable is in Hulliger (1991, Chapter 2).

Problems with nonuniqueness of the minimum or when the minimum is not attained on K are easily resolved in practice by inspection of the function $r(F_S, k)$. (If there are several global minima choose the one with smallest tuning constant to obtain more robustness.) The bias part of $r(F_S, k)$ involves the slope $\beta_{LS}(F_S)$ of the HT-estimator. By this term the sensitivity of the HT-estimator is transferred to MER-estimators and thus the robustness of RHT-estimators is lost again. But if the MER-estimator should be consistent for the population mean there is no way around a consistent and therefore nonrobust estimator in the bias part of the risk estimator. Nevertheless MER-estimators are quantitatively less sensitive to outliers and more efficient than the HT-estimator if outliers occur (see Section 5).

It is even possible to bound the influence of outliers on the MER-estimator for finite samples without losing its (asymptotic) consistency. This is achieved by downweighting

the bias part in the estimated risk of the HT-estimator in an appropriate way (MER2-estimators, (Hulliger 1991, Paragraph 2.4.1)).

MER-estimators may be more efficient than the HT-estimator because their bias is more than compensated by the variance reduction due to the downweighting of outliers. How much can be gained quantitatively is explored in Section 5.

5. A SMALL SIMULATION STUDY

Simulations with populations of size $N = 128$ and with samples of size $n = 16$ are presented here. The sample design in Dey and Srivastava (1987) is used (Note that there is a factor 2 missing in their formula (2.3)). Dey and Srivastava propose to form $m > n/2$ groups. The group totals $\sum_{j=1}^m x_{ij}$ ($j = 1, \dots, m$) must fulfill the inequality $\sum_{j=1}^m x_{ij} / \sum_{i=1}^n x_i > (n-2)/(n(m-1))$. Thus the group totals are allowed only little variability and the groups are difficult to form in particular for larger samples (Hulliger 1991, p. 179).

The x_i ($i = 1, \dots, N$) are independent realisations according to a 5%-scale contaminated exponential distribution with origin at 1, i.e., $(X_i - 1) \sim 0.95 \text{Exp}(1) + 0.05 \text{Exp}(3)$, where $\text{Exp}(\theta)$ denotes the exponential distribution function $1 - \exp(-x/\theta)$. The shift +1 is introduced to lower the probability of negative responses in the regression through the origin model with symmetric errors.

The first response $y_U^{(1)}$, with acronym GODA, is a realization of independent normal variables distributed as $Y_i \sim \mathcal{N}(100x_i, x_i^2)$. This is the model under which the HT-estimator is optimal (cf. Godambe 1955). The response $y_U^{(2)}$ (HTLS) is a realization of independent variables distributed as $Y_i \sim \mathcal{N}(2x_i, x_i/4)$. This is the ideal model that yields the HT-estimator as LS-estimator. A third response $y_U^{(3)}$ (HTG) is created by the model $Y_i \sim 0.95\mathcal{N}(2x_i, x_i/4) + 0.05\mathcal{N}(2x_i, 9x_i/4)$. The residual outliers have 3 times larger scale. The response $y_U^{(4)}$ (HTE) has asymmetric outliers which are not related to the x -variable. The bulk of the data (120 observations) stems from the distribution $Y_i \sim \mathcal{N}(2x_i, x_i/4)$ of $y_U^{(2)}$ but 8 randomly chosen observations stem from $\text{Exp}(2.5)$. The population $y_U^{(5)}$ (HMT) stems from a distribution with expectation $0.4 + 0.25x_i$ and has a Gamma distribution with variance proportional to $x_i^{3/2}$. Thus the variable y has the distribution proposed in Hansen, Madow and Tepping (1983, p. 781). Finally a population $y_U^{(6)}$ (HMTE) is generated with 120 observations from the same distribution as $y_U^{(5)}$ but with 8 randomly chosen observations from the distribution $\text{Exp}(2)$. The six populations above are chosen to be realistic. They all use the same population of x -values (see Figure 1).

The RHT estimator in the simulation uses

$$\eta(x'_i, r'_i) = w(x'_i, k_x) \psi_{\text{Hub}}(r'_i, k_r \text{med}_S | r'_i |),$$

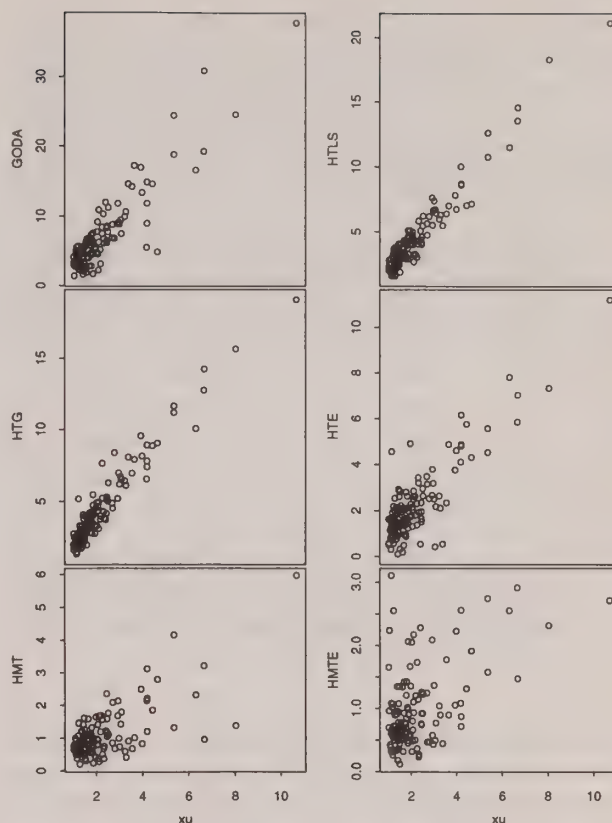


Figure 1. Populations of the Monte-Carlo Study.

with $w(x'_i, k_x) = \min(1, k_x \text{med}_U |x'_i| / |x'_i|)$ and $k_x = k_r = 2$. The weighting function $w(x'_i, k_x)$ corresponds to an asymmetric Huber-function $\psi_{\text{aHub}} = \min(x'_i, k_x)$, which downweights large x'_i only. The scale $\text{med}_S |r'_i|$ is the median of the absolute residuals evaluated at the solution of the preceding iteration of the iteratively reweighted least squares algorithm. The MER-estimator uses the same η with tuning constants k_x, k_r evaluated at 20 points which lie on the diagonal of the range of k_x and k_r . S-PLUS functions for the calculation of the estimators may be obtained from the author.

For each of the populations a set of 400 samples was drawn to evaluate the estimators. The obtained precision is sufficient to draw conclusions (see the standard errors of the efficiencies in Table 1).

The results are presented in Table 1. The relative bias of the RHT-estimator is always larger than the relative bias of the MER-estimator. The biases of the two estimators have the same sign, except when they are very small. With the exception of populations HTE and HMTE the variance of the RHT-estimator is larger than the variance of the MER-estimator. While the RHT-estimator loses 9% efficiency at population GODA, where the HT-estimator should be optimal, the MER-estimator loses little. With population HTLS, where the HT-estimator is the least squares estimator, the RHT-estimator loses about 12%.

Table 1

Monte-Carlo simulations with RHT- and MER-estimator

	Populations					
	GODA	HTLS	HTG	HTE	HMT	HMTE
MC-mean of HT	6.996	4.531	4.483	2.271	1.068	0.991
Rel. bias of RHT	-0.002	-0.001	-0.009	-0.009	0.006	-0.052
Rel. bias of MER	0.000	-0.001	-0.007	-0.008	-0.002	-0.035
Rel. SE of HT	0.067	0.041	0.044	0.098	0.107	0.170
Rel. SE of RHT	0.070	0.044	0.040	0.087	0.117	0.144
Rel. SE of MER	0.068	0.042	0.040	0.091	0.107	0.146
Eff. of RHT	0.911	0.876	1.110	1.310	0.827	1.234
Eff. of MER	0.969	0.981	1.158	1.194	0.989	1.284
MC-SE of eff. RHT	0.020	0.017	0.073	0.009	0.018	0.001
MC-SE of eff. MER	0.003	0.009	0.037	0.002	0.013	0.002

NOTE: Relative bias and relative standard error (rel. SE) are biases and standard errors divided by the MC-mean of the HT-estimator. Efficiencies (Eff.) are MSE of the HT-estimator divided by the MSE of the estimator. Estimated standard errors of these Monte-Carlo estimates of efficiency are given in the last two lines.

The efficiency loss of the MER-estimator is once again small. Population HTG contains symmetric residual outliers. The RHT-estimator gains about 11% (but see the error of 7.3%) and the MER-estimator about 16%. Under the asymmetric outliers of population HTE the gain of the RHT-estimator is 31% while the MER-estimator gains 19%. If neither the regression through the origin, nor the symmetry of errors or the proportionality of their variance to the explanatory variable holds, *i.e.*, for population HMT, then the RHT-estimator loses 17% compared with the HT-estimator while the MER-estimator loses practically nothing. If in such a population a few asymmetric outliers turn up like in population HMTE then both robust estimators gain considerably against the HT-estimator, namely 23% and 28% respectively.

In conclusion from this limited simulation the MER-estimator loses little in terms of MSE, compared with the HT-estimator, when there are no outliers in the population. It gains moderately in populations with symmetric outliers and considerably when the outliers are asymmetric. The RHT-estimator loses more under ideal situations than the MER-estimator. The adaptivity of the MER-estimators pays off.

Extensive simulations with infinite populations in Hulliger (1991) confirm these conclusions and show that the gains of robust estimators may be very large for skew populations with outliers. However the possible efficiency gains with robust estimators vanish for large samples since then the bias dominates MSE. On the other hand if the outliers that turn up in a sample are not representative, *e.g.*, if they are uncorrected coding errors, then the robust estimators are much more efficient than the HT-estimator for all sample sizes.

ACKNOWLEDGEMENTS

This article is an outgrowth of the authors Ph.D. thesis at ETH Zürich. The author would like to thank Prof. F.R. Hampel and Prof. H.R. Künsch, for their advice. The author is grateful for the valuable comments of two anonymous referees on drafts of this paper.

REFERENCES

- CHAMBERS, R.L. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association*, 81, 1063-1069.
- COCHRAN, W.G. (1977). *Sampling Techniques* (3rd. Ed.). New York: Wiley.
- DEY, A., and SRIVASTAVA, A.K. (1987). A sampling procedure with inclusion probabilities proportional to size. *Survey Methodology*, 13, 85-92.
- FULLER, W.A. (1991). Simple estimators of the mean of skewed populations. *Statistica Sinica*, 1, 137-158.
- GLASSER, G.J. (1962). On the complete coverage of large units in a statistical study. *International Statistical Review*, 30, 28-32.
- GODAMBE, V.P. (1955). A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society, Series B*, 17, 269-278.
- GWET, J.-P., and RIVEST, L.-P. (1992). Outlier resistant alternatives to the ratio estimator. *Journal of the American Statistical Association*, 87, 1174-1182.
- HAMPEL, F.R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69, 383-393.
- HAMPEL, F.R., RONCHETTI, E.M., ROUSSEEUW, P.J., and STAHEL, W.A. (1986). *Robust Statistics*. New York: Wiley.
- HANSEN, M.H., MADOW, W.G., and TEPPING, B.J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association*, 78, 776-807.
- HIDIROGLOU, M.A., and SRINATH, K.P. (1981). Some estimators of a population total from simple random samples containing large units. *Journal of the American Statistical Association*, 76, 690-695.
- HUBER, P.J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35, 73-101.
- HULLIGER, B. (1991). Nonparametric M-estimation of a population mean. Doctoral Dissertation ETH No. 9443, ETH Zürich.
- IACHAN, R. (1984). Sampling strategies, robustness and efficiency: the state of the art. *International Statistical Review*, 52, 209-218.
- JAECKEL, L.A. (1971). Robust estimates of location: symmetry and asymmetric contamination. *Annals of Mathematical Statistics*, 42, 1020-1034.

- KISH, L. (1965). *Survey Sampling*. New York: Wiley.
- LITTLE, R.J.A., and SMITH, Ph.J.(1987). Editing and imputation for quantitative survey data. *Journal of the American Statistical Association*, 82, 58-68.
- OEHLERT, G.W. (1985). The random average mode estimator. *Annals of Statistics*, 13, 1418-1431.
- RAO, J.N.K. (1966). Alternative estimators in PPS sampling for multiple characteristics. *Sankhyā A*, 28, 47-60.
- RIVEST, L.-P. (1993). Winsorization of survey data. *Proceedings of the 49th Session, International Statistical Institute*.
- SEARLS, D.T. (1966). An estimator for a population mean which reduces the effect of large observations. *Journal of the American Statistical Association*, 61, 1200-1204.
- SHOEMAKER, L.H., and ROSENBERGER, J.L. (1983). Moments and efficiency of the median and trimmed mean for finite populations. *Communications in Statistics, Simulations and Computations*, 12(4), 411-422.
- SMITH, T.M.F. (1987). Influential observations in survey sampling. *Journal of Applied Statistics*, 14, 143-152.
- STATISTICAL SCIENCES, INC. (1990). *S-PLUS Software*, Seattle: Statistical Science, Inc.

Visitor Sample Surveys

RONALDO IACHAN and SUZANNE S. KEMP¹

ABSTRACT

This paper discusses the design of visitor surveys. To illustrate, two recent surveys are described. The first is a survey of visitors to National Park Service areas nationwide throughout the year (1992). The second is a survey of recreational users of the three-river basin around Pittsburgh, Pennsylvania, during a twelve-month period. Both surveys involved sampling in time with temporal as well as spatial stratification. Sampling units had the form of site-period pairs for the stage before the final, visitor sampling stage. Random assignment of sample sites to periods permits the computation of unbiased estimates for the temporal strata (*e.g.*, monthly and seasonal estimates) as well as estimates for strata defined by region and by type of use.

KEY WORDS: Recreational user; Sampling in time; Site-period.

1. INTRODUCTION

Surveys of visitors present unique challenges that are rarely discussed in the statistical literature. This paper attempts to fill this gap by describing the design and emphasizing the common features of two surveys recently conducted by the Research Triangle Institute (RTI). We hope that the lessons learned in these efforts will be beneficial to researchers planning similar surveys.

The first survey was a study of visitors to National Park Service (NPS) areas jointly conducted for the National Park Service by RTI and HBRS, Inc. This study involved a probability sample of park visitors that represented visitors to 323 NPS areas nationwide (except Alaska) throughout the year (1992). We will refer to NPS areas as parks for simplicity while pointing out that the NPS areas include locations of historical and cultural parks. The main objective of the NPS study was to assess the visitors' experiences and problems with particular attention to those related to aircraft overflights (*e.g.*, noise and other possible annoyances). A variety of data were also collected in a mail survey for a subsample of selected visitors.

The second survey was a study of recreational users of the Pittsburgh-area three-river basin along the Monongahela, Allegheny and Ohio Rivers in 1992 (or more precisely, between February 1992 and January 1993). This survey was jointly conducted for the Ohio River Valley Sanitation Commission by RTI and Terrestrial Environmental Specialists. The study area included a 40-mile segment of the Ohio River, a 24-mile segment of Monongahela River and a 7-mile segment of the Allegheny River. The primary objective of the Three-River Study was to construct a baseline profile of recreationists in the area and to model the economic value they assign to various activities. Three basic types of recreational activities were distinguished: boating, fishing and park use.

The Three-River Study is the most comprehensive of a series of studies conducted by RTI to assess environmental impact in a number of states. These studies estimate possible reductions in economic or recreational value assigned by actual and potential recreational users to areas that have been or might be affected. While a wider survey of potential users of such areas may consider a telephone sample design, a visitor intercept survey design is found necessary to capture users at a point in time close to actual use.

A discussion of design issues in visitor surveys such as these has been recently provided in Kalton (1991) including issues related to sampling in time and space that are crucial in our framework. In its simplest form, a prototype, two-stage sample design for a visitor survey considers site-period pairs as primary sampling units (PSUs) from which visitors are selected in the second stage. Examples include exit polls (see, for example, Levy 1983), shopping mall intercept surveys (see, for example Sudman 1980) and other transportation and traffic surveys (Gough and Ghangurde 1977; Kish, Lovejoy and Rackow 1961). Among the design issues salient in visitor surveys, the following general problems may be singled out:

- It is desirable to select with greater probabilities those site-periods with larger numbers of visitors; stratification and PPS selection are then effective design features.
- Data collection arguments are key for the specification of the period length and of sampling rates within site-periods; *e.g.*, trade-offs occur between the potential for the field staff to be too busy (short periods, high sampling rates) or not busy enough (long periods, low sampling rates).
- Analytic objectives as well as efficiency suggest temporal stratification dimensions as season, month, weekend versus weekday, and even time-of-day; *e.g.*, the need for seasonal estimates suggests the use of seasons or months as strata for the selection of periods.

¹ Ronaldo Iachan, Research Statistician and Suzanne S. Kemp, Statistician, Research Triangle Institute, 3040 Cornwallis Road, Research Triangle Park, NC 27709-2194, U.S.A.

The two surveys discussed in this paper share the primary objective of characterizing the visitor population in the area or the nation over an entire year. They differ, however, in the priority estimates that lead to the basic design features in each case. In the Three-River Study but not in the NPS study, reasonably precise monthly estimates needed to be computed. For the former study, then, the temporal sampling units – days – were stratified into months. Spatial stratification of the Three-River Study sites was geographic and by recreation type (boating, fishing or park sites).

For the NPS study, primary stratification was by park type. Some park areas needed to be included with certainty into the sample to satisfy legislative requirements. In these and other selected parks, it was further desired to compute park-specific estimates. In these park areas, labelled intensive parks, we then decided to select relatively more site-periods. The initial design optimization problem was how to allocate the sample size to the sampling stages, *i.e.*, to decide how many parks and how many site-periods per park should be selected. Section 3 discusses a solution for this problem which is a function of the intraclass correlations within parks and within periods. The design optimization for the NPS survey also applied to the temporal and spatial strata at the intermediate sampling stages, between park areas at the first stage and site-periods at the penultimate stage (keeping in mind that visitors are selected at the final stage).

These two surveys illustrate issues such as temporal stratification, the choice of appropriate sampling units, and random assignment of spatial units to temporal units. Section 2 outlines the common aspects of the two studies as well as their basic differences. Sections 3 and 4 describe the design of the NPS Visitor Survey and of the Three-River Study, respectively. Section 5 discusses the weighting procedures used for the surveys. A brief overview and some conclusions are presented in Section 6.

2. OVERVIEW OF SAMPLE DESIGNS: PARALLELS AND CONTRASTS

For both surveys, the ultimate visitor samples were selected via intercept sampling as visitors left the sample locations at the selected time periods. Exit interviews were necessary to reflect their attitudes immediately following their recreational experiences. Also, in both studies, visitors were selected from sample site-period pairs. The use of site-period pairs as sampling units dates back to Kish, Lovejoy and Rackow (1961). This sampling unit definition permits the selection of visitors according to a data collection schedule that specifies which sites will be covered at which points in time. Unlike the Three-River Study, the selection of site periods was not the first stage

for the NPS Study. The primary sampling units (PSUs) for this study were NPS areas, or parks. The NPS survey involved several stages of selection described in the next section.

Additionally, both studies used temporal frames of days, and eligible data collection periods within days, to permit inferences about the entire year. The designs included the selection of time periods so that each eligible period has a known, positive probability of selection. Although both studies involved temporal frames, the structure of the frames and selection of days for each study were quite different.

For example, the sample for the Three-River Study was selected as twelve independent monthly samples. Each monthly sample has essentially the same, stratified random sampling design but a different sample allocation and different sample sizes were used in different months. This design took into account seasonal variations in recreational patterns, and enabled estimation for each month and stratum (*e.g.*, by type: boating, fishing or park). Both spatial and temporal frames were allowed to vary from month to month. The stratification and allocation for this sample are discussed in Section 4.

In contrast, the temporal frame for the NPS visitor survey first considered two-month blocks for each sample park (PSU). The use of two-month periods as (second-stage) sampling units in time efficiently met the survey objectives for two basic reasons. First, it allows the effective (geographic) concentration of staffing resources and staggered data collection throughout the year. Second, this choice of period permitted capturing seasonal fluctuations in park visitation across the park system, resulting from some parks having relatively higher visitation in the spring, others in the fall months, and so on.

One two-month block was selected for each sample park so that data collection could be effectively concentrated in time. Then, at the next stage of temporal selection, days were selected from within the two-month block for each sample park. Like the parks themselves at the first stage, these two-month blocks were selected with probabilities proportional to size (PPS), with the size measure being the aggregate visitation.

The sample sizes and allocation to the several sampling stages were carefully balanced to minimize clustering effects associated to clusters in time and space. For the Three-River Study, this clustering occurs at the first stage of selection where sampling units are sites and time periods. The allocation also considered the varying sample sizes used in successive, independent monthly samples. For the NPS survey, clusters in time were a result of the two-month blocks and sample days periods selected at different stages. Spatial clusters resulted from the use of parks and park exits as sampling units for this survey.

The next section describes in more detail the design of the NPS survey.

3. NPS SURVEY SAMPLE DESIGN

The design of NPS visitor survey capitalized on auxiliary information of various kinds and sources:

- Information obtained in previous studies (*e.g.*, park rankings based on noise exposure and NPS staff classifications).
- Information available in NPS data bases (*e.g.*, park visitation data by month).
- Information collected from NPS staff specifically for design purposes (*e.g.*, an inventory of the park exits for each sample park and number of vehicles leaving each of the exits).

The next subsections describe the various stages of selection for the visitor intercept survey. This survey component will also be designated the frontcountry survey to distinguish it from a survey of backcountry users that was conducted in tandem in the sample parks. Subsection 3.5 describes the backcountry survey as well as a mail survey administered to a subsample of frontcountry respondents and to sample backcountry users.

The fourth-stage selection for the mail survey involved additional stages (and phases) of selection. For the visitor intercept survey, groups (*i.e.*, vehicles) were selected as an ultimate cluster: all persons in a sample group were solicited for an interview. For the mail survey, groups were subsampled, and one person was subsampled from each subsampled group.

3.1 Frame Construction and First-Stage Sampling

We constructed a sampling frame for the selection of parks by compiling NPS information on park visitation (monthly and annual) and on noise exposure, information that was used in the sample design in two distinct ways: for stratification and for assigning size measures. The latter information was based on two different sources: (a) a previous NPS study which ranked parks according to potential exposure, and (b) a classification of parks performed independently by NPS staff (park superintendents, regional staff *etc.*).

In consultation with NPS, RTI combined these two classifications for noise exposure to construct eleven strata. The stratification partitioned parks into categories – very high, high, low and very low. Strata were divided into two substrata using the rankings in the stratum. (Note that the “medium” stratum was not subdivided due to its small park count.)

In addition to noise exposure, these strata incorporate three classes of parks that deserve separate treatment: (1) urban and suburban park areas, (2) parks with missing data on visitation (needed for PPS selection), and (3) parks whose elongated shapes present unique problems of access and reduce the meaning of prior exposure assessments. These classes were sampled at a much lower rate than the

other strata; the lowest sampling rate is in the urban stratum (1 in 79).

The certainty stratum included the seven parks that were mandated by legislation to be included in the study. In addition, it included those parks whose aggregate (annual) visitation rates were so large as to ensure selection into the first-stage sample. The 39 sample parks are listed in Exhibit 1; a 40th selected park (Grand Teton) was dropped from the sample for political reasons.

Design optimization calculations led to a first stage sample size of about 40 sample parks, yielding a total of 405 site-periods (or exit-days in this case) selected across intensive and non-intensive parks. As described in Section 3.3, 15 exit-days were selected in each of the three intensive parks, and 10 exit-days were selected in each of the 36 non-intensive parks in the sample. An accurate optimization would require variance components for the between-park and within-park variances. These variance components were approximated using data from a previous study as well as monthly visitation data available for all NPS areas.

3.2 Selection of Two-Month Blocks

While the selection of two-month blocks was with probability proportional to size (PPS), practical requirements were also taken into consideration.

First, training local park staff immediately before data collection was desired. Geographic clusters of parks were formed so that trainers could visit parks in one cluster in one trip. Specifically, the 39 sample parks were grouped into 14 such clusters. This requirement led to the selection of a two-month period for each park cluster. Thus, fourteen two-month blocks were selected, one for each cluster. The size measure used for each selection was the aggregate visitation over the parks in the cluster.

Exhibit 1 shows the sample parks in each cluster, and the sample two-month block selected for the cluster (*i.e.*, for every park in the cluster). Three strata – groups of clusters and hence of parks – were formed for the selection of two-month blocks:

- (a) In the “very-high summer” stratum-1 (with 3 clusters), the frame of two-month blocks contained only the summer-peak period, July-August, which was then selected with certainty for these clusters;
- (b) In the “high-summer” stratum-2 (with 7 clusters), the frame of two-month blocks contained 11 overlapping two-month blocks;
- (c) In the “low-summer” stratum-3 (with 4 clusters), the frame of two-month blocks contained 5 non-overlapping two-month blocks.

The rationale for this temporal stratification was two-fold: it ensured that the data collection was spread throughout the year, and it distinguished parks where a vast majority of the visitation occurs during the summer from those with a more uniform visitation pattern.

Exhibit 1**Park Clusters, Second-stage Strata and Selected Two-month Blocks**

	Selected Period
Stratum 1: Very High Summer*	
Cluster 2: Mount Rainier, N. Cascades, Olympic	July-August
Cluster 4: Glacier, Yellowstone	July-August
Cluster 10: Sleeping Bear, Perry's Victory	July-August
Stratum 2: High Summer	
Cluster 3: Lassen, Yosemite, Kings Canyon/Sequoia	August-September
Cluster 5: Dinosaur, Rocky Mtn., Mt. Rushmore	June-July
Cluster 6: Glen Canyon, Grand Canyon, Walnut Canyon	June-July
Cluster 8: Bandelier, Lake Meredith	May-June
Cluster 11: Cape Cod, Delaware Gap, Gettysburg	August-September
Cluster 12: Shenandoah, Fredericksburg, Assateague	June-July
Cluster 13: Great Smoky, Cape Hatteras, Fort Sumter	June-July
Stratum 3: Low Summer	
Cluster 1: Haleakala, Hawaii Volcanoes	March-April
Cluster 7: Lake Mead, Saguaro, Casa Grande	March-April
Cluster 9: Hot Springs, Wilson's Creek, Buffalo	October-November
Cluster 14: Cumberland Isd., Canaveral, Everglades, Gulf Island	March-April

* Certainty selection of July-August for each cluster in this stratum.

We selected two-month blocks with different procedures in the two strata as described below.

- (1) For park clusters in the former (high-summer) stratum, eleven overlapping periods were included in the temporal frame; January-February, February-March, . . . , November-December. One such period was then selected with probability proportional to size (PPS). Note that for this stratum each month was included in two frame periods except for January and December. The probability of selection for these two winter months was thus reduced even further (beyond the already small probability assigned to the winter periods with the PPS procedure).
- (2) For park clusters in the latter (low-summer) stratum, five non-overlapping two-month periods constituted the temporal frame: January-February, March-April, May-June, September-October, November-December. Note that the two-month summer period, July-August, was excluded from the frame for these parks to ensure the selection of other, non-summer months. For each cluster in this stratum, one of these five two-month periods was selected with PPS.

3.3 Third-stage Sampling

Sampling units at the third stage were exit-day pairs. The third-stage sampling is easier to envisage as the combination of two independent selections: (a) selecting days from the temporal window (2-month block) drawn at the second stage for the park, and (b) selecting exits from a list of exits specified for the park. A final step consisted of the random assignment of sample days to sample exits.

The sample of days was stratified by weekdays versus weekend days (including major holidays). Sample days were selected with equal probabilities within each of these two strata. The sample of exits was selected with probabilities proportional to size (PPS). The size measure assigned to each exit was the relative use of the exit among all the exits listed for the selected park. This usage measure was derived with the aid of local park staff.

The third-stage sample design distinguished two groups of parks designated as intensive (3 parks: Grand Canyon and the two Hawaii parks) and non-intensive (36 remaining parks in the sample). Sample sizes in non-intensive parks were 10 exits and 10 days, and hence 10 exit-day pairs. In the three intensive parks, 15 exits and 15 days were selected. Equal allocation to weekend/weekday strata was used in non-intensive parks: 5 weekend days and 5 weekdays were independently selected in each of these (36) sample parks. The allocation was approximately equal in intensive parks with the selection of 7 weekend days and 8 weekdays.

3.4 Fourth-stage Sampling

At the fourth stage, park visitors were intercepted in the selected exit-days. A systematic random sample of visitors or (visitor groups) exiting the park was selected with a fixed sampling interval for each selected third-stage unit (exit-day). The interval was allowed to vary from day to day to capitalize on the experience of previous days and on the variability in visitation across days and exits. Each visitor found in the selected eligible groups was screened for eligibility, and interviewed if eligible (adult visitor).

3.5 Backcountry and Mail Surveys

The same sample of parks (PSUs) was used for a mail survey of frontcountry and backcountry users in the two-month period selected for the park. The sample for the backcountry survey was restricted to the subset of sample parks with some backcountry use. Within each such sample park, the (third-stage) sampling frame for this survey component was based on backcountry permits issued during the data collection time window (two-month block) established for the park. The ultimate sample of backcountry users was then selected with equal probabilities from permit lists provided by park staff. A subsample of visitors selected in the intercept survey were also selected for the mail survey.

The mail survey sample was based on the same third-stage sample of exit-days selected for the intercept survey. For each selected exit-day, a fixed number of groups was subsampled from groups responding to the intercept survey (this number was 15 in intensive parks and 10 in non-intensive parks). A further stage of subsampling was that of one person from within the respondents in each group subsampled. This subsampling of groups and persons was with stratified random sampling to control the demographic composition of the final sample.

4. SAMPLE DESIGN FOR THREE-RIVER SURVEY

4.1 Frame Construction and Stratification

First, RTI and Terrestrial Environment Specialists (TES) constructed a sampling frame based on an inventory of all sites in the Three-River Study area. Then a stratified multistage sample was selected independently for each of the twelve months of the study. First-stage sampling units were site-periods, and second-stage units were individuals engaged in recreation in the selected site-periods. Temporal and spatial stratification were used for the first-stage sampling of time periods and sites.

Primary stratification along the temporal dimension was by month, and primary stratification of sites was by use type: sites (access points) were classified as boating, fishing, or parks. Each primary stratum was divided into six geographic areas, or pools, defined as the river areas between locks. The fishing stratum was further substratified by the presumed use intensity as low or high use: high-use sites are those below locks and dams.

Each monthly sample of site periods was selected with stratified random sampling. Advantages of selecting independent monthly samples included: monthly and seasonal estimates can be computed, and some design features may be changed from month to month.

For example, this design permitted altering the spatial frame from one month to the next with the addition or deletion of sites. In particular, the boating stratum for the winter months (November through April) was restricted to boat ramps open that season. Further, several fishing sites included in the frame for the first few months were found inaccessible and deleted from the frame for the subsequent monthly samples.

Other design features that changed in successive months include: the second-stage sampling rates for selecting eligible users, and the data collection windows used in the morning, afternoon and evening periods for sites of each type. Varying data collection windows were used in different months of the study. These periods were defined using sunrise and sunset information as well as expected patterns of use of the various types. The winter months (November-April) included two periods per day while the summer months (May-October) included three periods per day.

The temporal (sub)stratification of each monthly sample was by weekend days versus weekdays. It is worth noting that the weekend stratum also included major holidays.

4.2 Sample Selection

As noted above, we selected an independent first-stage sample of site-periods for each of the 12 months of the study. Each monthly sample had two components: (a) a stratified random sample of “*n*” sites, and (b) a stratified random sample of “*n*” periods.

Following selection of the sample periods for each month, the sample sites were randomly assigned to the selected periods. The assignment of sites to periods was entirely at random for the months of February through June but was modified in subsequent months. From July on, a sample of time periods was independently selected for each type stratum with the random assignment taking place within stratum. The allocation of the sample time periods (e.g., the number of morning periods and the number of evening periods included in the sample) varied from stratum to stratum. With this more flexible method, relatively more fishing sites could be assigned to morning period and more boating sites to afternoon periods, for example. Exhibit 2 shows the sample sizes – sites and periods – used in random assignment each month.

Exhibit 2

Sample Sizes Used in Random Assignment of Sites to Periods for each Monthly Sample of Three-River Study

Month	Overall*	Sample Size		Parks	Marinas	Regattas	
		Boating	Fishing			Boating	Marinas
1		8	12	4			
2	20						
3	28						
4	28						
5	36						
6	36						
7		17	22	6	12		
8		10	22	6	6	6+	6+
9		10	14	6	6		
10		10	12	4			
11		8	12	4			
12		8	12	4			

* For these (5) months, the assignment took place for the entire collection of sample sites and periods (NOT blocked by site type). For the remaining months the assignment was within each type stratum, a process which involved the selection of independent samples of time periods for each type.

+ In August, the assignment of regatta sites to periods was performed first, separately. Following the assignment, the sample site-periods associated with regattas were shifted either to the boating or to the marina strata depending on the site type.

Second-stage sampling rates were specified for each of the three primary strata prior to each month of data collection. These stratum-specific rates were determined based on the experience of the previous months, and were distributed to the field interviewers along with the month's data collection schedule. Individuals were selected with systematic random sampling within each site-period.

4.3 Marina Survey and Special Events

A marina survey was conducted in the months of June to September. A sampling frame of 48 marinas was based on the TES inventory that was updated in late May 1992.

The sampling method for the phase-in month of June differed slightly from that used for the subsequent monthly samples for marina sites. The June sample was a supplement of 12 marina sites coupled with a sample of 12 days. The marina samples for the months of July to September were selected considering the marina frame as a fourth stratum. The selection procedures were then similar to those in the other three type strata; specifically, (a) a sample of “*n*” marina sites was selected, (b) a sample of “*n*” time periods was selected, and (c) the sample sites were randomly assigned to the selected periods.

It is worth pointing out that some of the marina sites were also included in the boating stratum. In such cases, two distinct frame units were created for the same site. This situation also arose for some sites used for both boating and fishing, and such sites were included in both strata.

In addition to the marina survey, we identified special events taking place in the study area over the 12-month study period. Most of these events were handled in a way similar to weekends and holidays by assigning them to a stratum to be oversampled. A special category of interest was comprised of the regattas occurring in the summer months. For two monthly samples (July and August), we identified the regatta dates as well as the sites affected by each regatta. We then constructed a separate (fifth) stratum to include these site-periods. The first-stage sample allocation to the regatta stratum reflected the oversampling desired for this stratum. As shown in Exhibit 2, the sampling procedure used to select site-periods (first-stage units) from the regatta stratum also differed to that used in the other four strata. Sample site-periods were directly selected in one step from the subset of site-periods in the stratum, *i.e.*, no random assignment was needed.

5. SURVEY WEIGHTING

5.1 NPS Survey Weighting

Sampling weights were first computed for each of the first three stages of selection. The first-stage sampling weight for each sample park was the reciprocal of the selection probability for the park. The second-stage sampling weight for each sample two-month block was similarly computed. The sum of the first-stage sampling weights overall (or in a stratum) was the number of parks in the frame (or in a stratum).

Third-stage weights for sample exit-days were the product of two factors associated with the selection of days and exits. Note that for each selected park and two-month block, the sum of the former set of weights in a temporal stratum

(weekend vs. weekdays) is the number of days in the stratum, and the sum of the latter set of weights is the number of exits listed in the park. These weights were adjusted for nonresponse which arose at the third stage because in a few parks, data collection did not take place in some selected exit-days. In a given park with this data collection shortcoming, the sum of the adjusted third-stage weights over the active exit-days was made equal to the sum of the sampling weights over all selected exit-days in the park.

Fourth-stage weights were computed at a group-level and at a person-level. Group-level weights are assigned to all participating groups in a sample exit-day, and have the same value for the groups in the same exit-day. Similarly, person-level weights are assigned to all persons intercepted in a sample exit-day. The fourth-stage sampling weights were computed as the reciprocal of the sampling rate specified for the sample exit-day. These weights were then adjusted for group and person-level nonresponse.

The mail survey sample was based on the same third-stage sample of exit-days selected for the intercept survey. For each selected exit-day in non-intensive parks, 10 groups were first selected with equal probabilities from among the participating groups; then, one person was subsampled from all intercept survey respondents in each selected group. A similar procedure was used in intensive parks with the exception that the number of selected groups per exit-day was 15 rather than 10.

The sampling weight for each mail survey record is the product of $WTB = \text{number of intercept respondents in the group}$, and $WTA = \text{number of participating groups}/10$ [non-intensive parks]. For intensive parks, the denominator of WTA is 15 rather than 10.

These weights were adjusted for mail survey nonresponse using exit-days (within park) as weighting classes. Thus, the sum of the adjusted weights for all respondents coming from the same exit-day is the same as the sum of the (unadjusted) weights for all persons intercepted in the exit-day.

For the backcountry survey, the sampling frame for each eligible park (in the subset of sample parks with backcountry use) was based on lists of permits issued during the data collection period: eligible permits were associated with exit dates in this period. A sample of 5 groups per day was selected within each park from the set of permits linked to that day. One person was subsampled from each group. The weight computation parallels that for the mail survey with

$$WTBACK = BACKA * BACKB,$$

where for each day:

$$BACKA = (\text{number of groups linked to the day})/5$$

$$BACKB = \text{number of persons in group}.$$

Analysis weights for the backcountry survey resulted from nonresponse adjustments made using days as weighting classes.

5.2 Three-River Survey Weighting

As each monthly sample of site-period units was selected, we computed sampling weights that reflected the selection probabilities for the site-period pairs. Initial weights were the product of two sets of weights computed for each monthly sample: (a) weights assigned to each site in the stratified random sample of sites of the given type (boating, fishing, marina and park), and (b) weights assigned to each period in the stratified random sample of periods. These weights were then inflated to take into account the random assignment of sample sites to periods (or vice-versa). Thus for each month, the sum of the site-period weights was equal to the number of site-period combinations in the frame.

The weight adjustment process started with the sampling weights associated with the selected site-days. An initial adjustment was made to the first-stage weights to account for site-periods that were found ineligible or that had missing sampling forms (*e.g.*, not sent by field staff). For this first-stage adjustment, we used the type-by-month strata as weighting classes. That is, the sum of the adjusted weights over the reduced set was made equal to the sum of the unadjusted weights over the entire sample within the type-by-month class. A final adjustment was made at the respondent-level to reflect (a) the systematic sampling interval used within site-periods, and (b) the survey non-response at the individual level.

As part of the weight check procedures, we computed the sum of the final analysis weights over the entire file, and also by month and by type of site. The weight sum should approximate the estimated total number of recreational users leaving the inventory sites during the data collection time window for each month and type of site.

6. CONCLUSION

This paper described the design of two surveys of recreational users that share a number of useful features. The sample designs include sampling in time as well as in space; site-periods are selected at the stage prior to sampling visitors. A spatial frame is then constructed side by side with a temporal frame. After sites are selected from the former and periods are selected from the latter frame, sample sites are randomly assigned to sample periods (or conversely). Sampling weights need to take into account this additional step of randomization. The findings of the NPS visitor survey are described in the study final report (National Park Service 1994). This analysis included a variety of regression models that investigate the impact of hearing and seeing aircraft flying over NPS areas.

Temporally, both studies represent periods throughout the year, that is, a user will have a positive probability of selection for any time of the year. Both studies also include

temporal stratification to reflect patterns of use and increase sampling efficiency.

Spatially, while the sample for the National Park Service study was a national sample of visitors to NPS areas, the sample for the Three-River Study was more restricted in spatial scope. Both studies, however, distinguished users of different types. For the NPS study, backcountry and frontcountry users were selected in two separate (third-stage) strata at the point where the two components branch out with the selection of permits for the backcountry survey. For the Three-River Study, visitors were classified by primary fishing, boating, or park use, and sites were stratified in a similar way. This design permitted the computation of precise estimates by type and by season.

The ultimate sampling unit for a survey of recreational users is the specific visit; thus, visitors may have multiple chances of inclusion in the sample to the extent that they use the target areas multiple times. It is worth noting that this structure is consistent with the objectives of such surveys.

Sampling weights accounted for the selection of time and space units at each stage and also for the random assignment step. The samples were designed to minimize the effects of unequal weighting on survey variances. The potential for severe unequal weighting effects was considered in combining different survey components. Examples in the two surveys include combining:

- (a) Backcountry and frontcountry components of NPS mail survey, and
- (b) Fishing, boating and park users in the Three-River Study.

Some disadvantages of this type of study design should also be pointed out. While sampling in time and random assignment introduce an element of statistical rigor and extend the range of valid statistical inferences, the methodology may be disrupted if field interviewers change the date assigned for a sample location because access may be difficult in the specified period or for other reasons. Noteworthy examples in the NPS study included hurricanes, park closed due to fugitives from justice, space shuttle launches, and severe snowstorms. While some of these occurrences may be minimized with the temporal stratification and allocation, others are clearly beyond the control of the statistician. However, the sampling statistician should be involved in interviewer training to stress that modifications in the sample schedule should be avoided at all costs, and should monitor any changes that do occur.

ACKNOWLEDGEMENTS

The authors are grateful to the Associate Editor and the referees for comments that helped improve the paper's presentation.

REFERENCES

- GOUGH, J.H., and GHANGURDE, P.D. (1977). Survey of Canadian residents returning by land. *Survey Methodology*, 3, 215-231.
- KALTON, G. (1991). Sampling flows of mobile human populations. *Survey Methodology*, 17, 183-194.
- KISH, L., LOVEJOY, W., and RACKOW, P. (1961). A Multi-stage probability sample for traffic surveys. *Proceedings of the Social Statistics Section, American Statistical Association*, 227-230.
- LEVY, M.R. (1983). The methodology and performance of election day polls. *Public Opinion Quarterly*, 54-67.
- NATIONAL PARK SERVICE (1994). Survey of Visitors to National Park Service Areas: Survey Findings. Report 94-2, No. 290940.12.
- SUDMAN, S. (1980). Improving the quality of shopping center sampling. *Journal of Marketing Research*, 17, 423-431.

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue (Vol. 19, No. 1 and onward) of *Survey Methodology* as a guide and note particularly the following points:

1. Layout

- 1.1 Manuscripts should be typed on white bond paper of standard size ($8\frac{1}{2} \times 11$ inch), one side only, entirely double spaced with margins of at least $1\frac{1}{2}$ inches on all sides.
- 1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.
- 1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.
- 1.4 Acknowledgements should appear at the end of the text.
- 1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

- 3.1 Avoid footnotes, abbreviations, and acronyms.
- 3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as “exp(·)” and “log(·)”, etc.
- 3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
- 3.4 Write fractions in the text using a solidus.
- 3.5 Distinguish between ambiguous characters, (e.g., w, ω ; o, O; l, 1).
- 3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

4. Figures and Tables

- 4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.
- 4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

5. References

- 5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, p. 164).
- 5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Avant de dactylographier votre texte pour le soumettre, prière d'examiner un numéro récent de *Techniques d'enquête* (à partir du vol. 19, n° 1) et de noter les points suivants:

1. **Présentation**
 - 1.1 Les textes doivent être dactylographiés sur un papier blanc de format standard (8 1/2 par 11 pouces), sur une face seulement, à double interligne partout et avec des marges d'au moins 1 1/2 pouce tout autour.
 - 1.2 Les textes doivent être divisés en sections numérotées portant des titres appropriés.
 - 1.3 Le nom et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.
 - 1.4 Les remerciements doivent paraître à la fin du texte.
 - 1.5 Toute annexe doit suivre les remerciements mais précéder la bibliographie.

2. **Résumé**

Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.

3. **Rédaction**
 - 3.1 Éviter les notes au bas des pages, les abréviations et les sigles.
 - 3.2 Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme exp(-) et log(-) etc.
 - 3.3 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin.
 - 3.4 Écrire les fractions dans le texte à l'aide d'une barre oblique.
 - 3.5 Distinguer clairement les caractères ambigus (comme w, ω; o, O; 1, I).
 - 3.6 Les caractères italiques sont utilisés pour faire ressortir des mots. Indiquer ce qui doit être imprimé en italique en le soulignant dans le texte.

4. **Figures et tableaux**
 - 4.1 Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).
 - 4.2 Ils doivent paraître sur des pages séparées et porter une indication de l'endroit où ils doivent figurer dans le texte. (Normalement, ils doivent être insérés près du passage qui y fait référence pour la première fois.)

5. **Bibliographie**
 - 5.1 Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence.
 - 5.2 La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.

REMERCIEMENTS

Les auteurs remercient le rédacteur associé ainsi que les arbitres pour leurs commentaires, qui ont aidé à améliorer la présentation de l'article.

BIBLIOGRAPHIE

- GOUGH, J.H., et GHANGURDE, P.D. (1977). Survey of Canadian residents returning by land. *Techniques d'enquête*, 3, 215-231.
- KALTON, G. (1991). L'échantillonnage des flux de populations humaines mobiles. *Techniques d'enquête*, 17, 197-210.
- KISH, L., LOVEJOY, W., et RACKOW, P. (1961). A Multistage probability sample for traffic surveys. *Proceedings of the Social Statistics Section, American Statistical Association*, 227-230.
- LEVY, M.R. (1983). The methodology and performance of election day polls. *Public Opinion Quarterly*, 54-67.
- NATIONAL PARK SERVICE (1994). Survey of Visitors to National Park Service Areas: Survey Findings. Report 94-2, No. 290940.12.
- SUDMAN, S. (1980). Improving the quality of shopping center sampling. *Journal of Marketing Research*, 17, 423-431.

compte de l'importance possible de ces effets lorsqu'on a combiné différentes composantes de l'enquête. Voici des exemples pour les deux enquêtes, soit la combinaison: a) des composantes arrière-pays et avant-pays de l'enquête postale du NPS et b) des personnes qui pratiquent la pêche sportive, la navigation de plaisance et qui fréquentent des parcs dans l'enquête portant sur les trois rivières.

Il faudrait aussi signaler certains inconvénients de ce type de plan d'échantillonnage. Alors que l'échantillonnage dans le temps et l'attribution aléatoire introduisent un élément de rigueur statistique et étendent la gamme d'inférences statistiques valides, la méthodologie peut être perturbée si les intervieweurs sur le terrain changent la date qui correspond à un emplacement échantillonné parce que l'accès à cet emplacement peut être difficile pendant la période précisée ou pour d'autres raisons. À titre d'exemples notables de situations de ce genre que l'on a rencontrées au cours de l'enquête du NPS, mentionnons des ouragans, des parcs fermés à cause de la présence de fugitifs recherchés par la justice, des lancements de navettes spatiales et de graves tempêtes de neige. Alors qu'on peut minimiser certaines de ces situations au moyen de la stratification temporelle et de la répartition de l'échantillon, le statisticien ne peut manifestement rien faire pour d'autres. Le statisticien qui s'occupe d'échantillonnage devrait participer à la formation des intervieweurs afin d'insister sur le fait qu'il faut éviter, à tout prix, de modifier le plan d'échantillonnage et il devrait surveiller toutes les modifications qui sont apportées.

6. CONCLUSION

Dans cet article, nous avons décrit le plan d'échantillon-

nage utilisé pour deux enquêtes, menées auprès de personnes qui utilisent des installations de loisirs et qui partagent un certain nombre de caractéristiques. Les plans d'échantil-
lonnage comprennent l'échantillonnage dans le temps ainsi
que dans l'espace; des paires site-période sont choisies au
degré qui précède l'échantillonnage des visiteurs. Nous
construisons alors une base de sondage spatiale en même
temps qu'une base de sondage temporelle. Une fois que
les sites sont choisis à partir de la première base de sondage
et les périodes, à partir de la dernière, les sites échan-
tillonnés sont attribués aléatoirement aux périodes échan-
tillonnées (ou inversement). Les poids d'échantillonnage
doivent tenir compte de cette étape additionnelle de rando-
misation. Les conclusions de l'enquête auprès des visiteurs
effectuée pour le NPS sont décrites dans le rapport final
de l'enquête (National Park Service 1994). Cette analyse
incluait divers modèles de régression employés pour évaluer
l'incidence liée au fait d'entendre et de voir des avions
survoler les zones administrées par le NPS.

Du point de vue temporel, les deux enquêtes portent sur
différentes périodes correspondant à toute l'année, c'est-
à-dire qu'un utilisateur aura une probabilité positive d'être
choisi pour toute période de l'année. Les deux études
comprennent aussi une stratification temporelle pour
refléter les habitudes d'utilisation et pour augmenter l'effi-
cacité de l'échantillonnage.

Du point de vue spatial, alors que l'échantillon utilisé
dans le cadre de l'enquête menée pour le National Park
Service était un échantillon national de visiteurs des zones
administrées par le NPS, l'échantillon employé pour
l'enquête portant sur les trois rivières englobait une région
plus limitée. Toutefois, dans les deux enquêtes, on faisait
une distinction entre différents types d'utilisateurs. Pour
l'enquête du NPS, les utilisateurs des installations de
l'arrière-pays et ceux des installations de l'avant-pays
étaient choisis (au troisième degré) dans deux strates
distinctes, les détenteurs de permis étant visés par l'enquête
portant sur l'utilisation des installations de l'arrière-pays.

Dans le cas de l'enquête portant sur les trois rivières, les
visiteurs étaient classés selon qu'ils s'intéressaient surtout
à la pêche sportive ou à la navigation de plaisance ou qu'ils
fréquentaient davantage les parcs, et les sites étaient strati-
fiés de la même façon. Ce plan d'échantillonnage a permis
le calcul d'estimations précises selon le type et la saison.
La dernière unité d'échantillonnage pour l'enquête sur
les personnes qui utilisent des installations de loisirs est la
visite; ainsi, les visiteurs peuvent avoir plusieurs chances
d'être inclus dans l'échantillon dans la mesure où ils utili-
sent plusieurs fois les zones visées par l'enquête. Il vaut
la peine de souligner que cette structure est compatible
avec les objectifs d'enquêtes de ce genre.

Les poids d'échantillonnage tenaient compte de la sélec-
tion des unités temporelles et spatiales à chaque degré ainsi
que de l'opération d'attribution aléatoire. Les échantillons
étaient conçus de façon à minimiser les effets de la pondé-
ration inégale sur les variances de l'enquête. On a tenu

Un sous-échantillon d'une personne a été choisi dans
chaque groupe. Le calcul des poids équivalait à celui qui est
fait pour l'enquête postale avec

$$WTBACK = BACKA * BACKB,$$

où, pour chaque jour:

$$BACKA = (\text{nombre de groupes liés au jour})/5$$

$$BACKB = \text{nombre de personnes dans le groupe.}$$

Les poids d'analyse pour l'enquête portant sur l'utili-
sation des installations de l'arrière-pays découlaient des
rajustements pour la non-réponse apportés en utilisant les
jours comme classes de pondération.

5.2 Pondération pour l'enquête portant sur les trois rivières

Au moment de la sélection de chaque échantillon mensuel
de paires site-période, nous avons calculé des poids d'échan-
tillonnage qui reflétaient la probabilité de sélection des
paires site-période. Les poids initiaux étaient le produit de
deux ensembles de poids calculés pour chaque échantillon
mensuel: a) Les poids attribués à chaque site dans l'échan-
tillon aléatoire stratifié de sites du type donné (navigation
de plaisance, pêche sportive, marina et fréquentation des
parcs) et b) les poids attribués à chaque période dans
l'échantillon aléatoire stratifié des périodes. On a alors
extrapolé ces poids pour tenir compte de l'attribution aléa-
toire des sites échantillonnés aux périodes (ou inversement).
Ainsi, pour chaque mois, la somme des poids pour les
paires site-période était égale au nombre de combinaisons
de paires site-période dans la base de sondage.

Le processus de rajustement des poids a commencé avec
les poids d'échantillonnage associés aux paires site-période
sélectionnées. Un rajustement initial a été apporté aux poids,
pour le premier degré, afin de tenir compte des paires site-
période qui avaient été trouvées inadmissibles ou pour
lesquelles des questionnaires manquaient (p. ex., question-
naires qui n'avaient pas été expédiés par les employés
travaillant sur le terrain). Pour ce rajustement, au premier
degré, nous avons utilisé les strates type par mois comme
classes de pondération. C'est-à-dire que nous avons rendu
la somme des poids rajustés sur l'ensemble réduit égale à
la somme des poids non rajustés sur tout l'échantillon dans
la classe type par mois. Un rajustement final a été apporté
au niveau du répondant pour refléter a) le pas de sondage
systématique utilisé dans les paires site-période et b) la non-
réponse à l'enquête au niveau des particuliers.

Dans le cadre des procédures de vérification de la pon-
dération, nous avons calculé la somme des poids d'analyse
finaux pour tout le fichier et aussi par mois et par type de
site. La somme des poids devrait correspondre, approxi-
mativement, au nombre total estimé de personnes utilisant
des installations de loisirs qui quittent les sites relevés
pendant la période de collecte des données pour chaque
mois et pour chaque type de site.

à jours de semaine) est le nombre de jours dans la strate, et la somme du dernier ensemble de poids est le nombre de sorties relevées pour le parc. Ces poids ont été rajustés afin de tenir compte de la non-réponse qui s'est produite au troisième degré parce que, dans quelques parcs, la collecte des données n'a pas eu lieu pour certaines des paires sortie-jour sélectionnées. Dans tout parc où l'on a observé ce défaut au niveau de la collecte des données, on a rendu la somme des poids rajustés, pour le troisième degré, pour les paires sortie-jour actives égale à la somme des poids d'échantillonnage pour toutes les paires sortie-jour sélectionnées dans le parc.

Pour le quatrième degré, les poids ont été calculés au niveau d'un groupe et au niveau d'une personne. Les poids au niveau d'un groupe sont attribués à tous les groupes participants de paires sortie-jour dans l'échantillon et ils ont la même valeur pour les groupes dans la même paire sortie-jour. De même, les poids au niveau de la personne ont été attribués à toutes les personnes interviewées sur le vif dans une paire sortie-jour dans l'échantillon. Les poids d'échantillonnage, pour le quatrième degré, ont été calculés comme l'inverse du taux de sondage précisé pour la paire sortie-jour dans l'échantillon. Ces poids ont alors été rajustés afin de tenir compte de la non-réponse au niveau du groupe et de la personne.

L'échantillon de l'enquête postale était basé sur l'échantillon, pour le troisième degré, de paires sortie-jour choisies pour les interviews sur le vif. Pour chaque paire sortie-jour sélectionnée dans un parc peu achalandé, on a tout d'abord choisi, avec probabilités égales, dix des groupes participants; puis, un sous-échantillon d'une personne a été choisi, parmi tous les répondants aux interviews sur le vif, dans chaque groupe sélectionné. Une procédure semblable a été utilisée dans les parcs très achalandés, sauf que, dans ce cas, on a choisi 15 groupes plutôt que 10 par paire sortie-jour.

Le poids d'échantillonnage pour chaque enregistré de l'enquête postale est le produit de WTB = nombre de répondants aux interviews sur le vif dans le groupe et de WTA = nombre de groupes participants/10 [pour les parcs peu achalandés]. Dans le cas des parcs très achalandés, le dénominateur de WTA est 15 plutôt que 10. Ces poids ont été rajustés, pour tenir compte de la non-réponse lors de l'enquête postale, à l'aide de paires sortie-jour (dans un parc) utilisées comme classes de pondération. Ainsi, la somme des poids rajustés pour tous les répondants provenant de la même paire sortie-jour est identique à la somme des poids (non rajustés) pour toutes les personnes ayant répondu aux interviews sur le vif dans la paire sortie-jour.

Pour l'enquête portant sur l'utilisation des installations de l'arrière-pays, la base de sondage pour chaque parc admissible (dans le sous-ensemble des parcs échantillonnés avec utilisation d'installations de l'arrière-pays) était fondée sur les listes de permis délivrés pendant la période de collecte des données: les permis admissibles étaient associés à des dates de sortie pendant cette période. Un échantillon de cinq groupes par jour a été choisi dans chaque parc à partir de l'ensemble des permis liés à ce jour.

des marinas comme une quatrième strate. Les procédures de sélection étaient alors semblables à celles utilisées dans les trois autres types de strate; plus précisément, a) un échantillon de "n" marinas a été choisi; b) un échantillon de "n" périodes a été choisi et c) les sites échantillonnés ont été attribués aléatoirement aux périodes choisies.

Il faut mentionner que certaines des marinas étaient aussi incluses dans la strate de la navigation de plaisance. Dans de tels cas, deux unités distinctes de la base de sondage ont été créées pour le même site. Cette situation s'est aussi produite pour certains sites utilisés à la fois pour la navigation de plaisance et pour la pêche sportive, et les sites de ce genre ont été inclus dans les deux strates.

En plus de l'enquête sur les marinas, nous avons relevé des événements spéciaux qui ont eu lieu dans la région et au cours de la période de douze mois visée par l'enquête. La plupart de ces événements ont été traités d'une façon qui ressemble à celle utilisée pour les jours de fin de semaine et les jours fériés en les attribuant à une strate qui sera suréchantillonnée. Une catégorie qui présentait un intérêt spécial était celle des régates, qui ont lieu pendant les mois d'été. Pour deux échantillons mensuels (juillet et août), nous avons relevé les dates des régates ainsi que les sites où chacune d'entre elles allaient se dérouler. Nous avons alors construit une (cinquième) strate distincte pour inclure ces paires site-période. Au cours du premier degré de l'échantillonnage, la répartition de l'échantillon dans la strate des régates reflétait le suréchantillonnage souhaité pour cette strate. Comme on le voit dans le document 2, la procédure d'échantillonnage utilisée pour choisir les paires site-période (unités du premier degré) à partir de la strate des régates diffèrait aussi de celle employée dans les quatre autres strates. Des paires site-période ont été choisies directement en une étape depuis le sous-ensemble des paires site-période dans la strate, c.-à-d. qu'aucune attribution aléatoire n'a été requise.

5. PONDERATION POUR LES ENQUÊTES

5.1 Pondération pour l'enquête du NPS

Les poids d'échantillonnage ont tout d'abord été calculés pour chacune des trois premières degrés de sélection. Pour le premier degré, le poids d'échantillonnage pour chaque parc dans l'échantillon était l'inverse de la probabilité de sélection du parc. Pour le deuxième degré, le poids d'échantillonnage pour chaque bloc de deux mois dans l'échantillon était calculé de la même façon. La somme globale (ou dans une strate) des poids d'échantillonnage pour le premier degré était le nombre de parcs dans la base de sondage (ou dans la strate).

Pour le troisième degré, les poids pour les paires sortie-jour dans l'échantillon étaient le produit de deux facteurs associés au choix des jours et des sorties. Il faut remarquer que pour chaque parc et pour chaque bloc de deux mois sélectionnés, la somme du premier ensemble de poids dans une strate temporelle (jours de fin de semaine par opposition

l'échantillon) variait d'une strate à l'autre. Avec cette méthode plus flexible, on a pu, par exemple, attribuer un nombre relativement plus élevé de sites de pêche sportive aux périodes du matin et plus de sites de navigation de plaisance aux périodes de l'après-midi. Le document 2 présente la taille des échantillons, sites et périodes, utilisés dans l'attribution aléatoire chaque mois.

Document 2

Taille des échantillons utilisés pour l'attribution aléatoire des sites aux périodes pour chaque échantillon mensuel de l'enquête portant sur les trois rivières				
Mois	Globale*	Taille de l'échantillon		Régates
		Nav- gation de pêche	Parcs Marinas	Nav- gation de plaisance Marinas
1	20	8	12	4
2	28			
3	28			
4	28			
5	36			
6	36			
7	17	22	6	12
8	10	22	6	6
9	10	14	6	6
10	10	12	4	4
11	8	12	4	4
12	8	12	4	4

* Pour ces cinq mois, l'attribution a eu lieu pour l'ensemble des sites et des périodes échantillonnés (il n'y a pas eu de regroupement selon le type de site). Pour les autres mois, l'attribution a été faite dans la strate de chaque type. Pour les autres mois, l'attribution a été faite dans la strate de chaque type. Pour aabr, on a tout d'abord attribué séparément les sites des régates aux périodes. Puis, les paires site-période échantillonnées associées aux régates ont été déplacées soit vers la strate de la navigation de plaisance, soit vers celle des marinas, selon le type de site.

Les taux d'échantillonnage, du deuxième degré, étaient précisés pour chacune des trois strates primaires avant la collecte des données pour chaque mois. Ces taux propres à la strate étaient déterminés d'après l'expérience acquise au cours des mois antérieurs, et ils étaient fournis aux intervieweurs sur le terrain avec le calendrier de collecte des données du mois. Les personnes ont été choisies par échantillonnage aléatoire systématique dans chaque paire site-période.

4.3 Enquête sur les marinas et les événements spéciaux

Une enquête sur les marinas a été réalisée de juin à septembre. On a utilisé une base de sondage composée de 48 marinas fondée sur le relevé effectué par TBS, qui a été mis à jour à la fin de mai 1992. La méthode d'échantillonnage utilisée pour le mois d'introduction de l'enquête sur les marinas, juin diffèrait légèrement de celle qui a été utilisée pour les échantillons des mois suivants. L'échantillon de juin était un supplément de 12 marinas combiné à un échantillon de 12 jours. Les échantillons de marinas pour les mois de juillet à septembre ont été choisis en utilisant la base de sondage

de plaisance, sites de pêche sportive ou parcs. Chaque strate primaire a été divisée en six régions géographiques, ou groupes, définies comme étant des segments de rivière entre des écluses. La strate de la pêche sportive a ensuite fait l'objet d'une sous-stratification en fonction de l'intensité d'utilisation prévue, soit intensive faible ou intensive élevée: les sites à intensive élevée étant ceux en aval des écluses et des barrages.

Chaque échantillon mensuel de paires site-période a été choisi par échantillonnage aléatoire stratifié. Les avantages de la sélection d'échantillons mensuels indépendants sont: la possibilité de calculer des estimations mensuelles et saisonnières et la possibilité de modifier certaines caractéristiques du plan d'un mois à l'autre.

Par exemple, ce plan a permis de modifier la base de sondage spatiale d'un mois à l'autre par l'ajout ou la suppression de sites. En particulier, la strate de la navigation de plaisance pour les mois d'hiver (de novembre à avril) a été limitée aux rampes de mise à l'eau qui étaient ouvertes pendant ces mois. De plus, plusieurs sites de pêche sportive inclus dans la base de sondage pour les premiers mois se sont révélés inaccessibles et ont été supprimés de la base de sondage pour les échantillons des mois ultérieurs.

Parmi les autres caractéristiques du plan qui ont changé au cours des mois, mentionnons: les taux d'échantillonnage, au deuxième degré, pour la sélection des utilisateurs admissibles ainsi que les périodes de collecte des données le matin, l'après-midi et en soirée pour les sites de chaque type. Des périodes de collecte des données variables ont été utilisées au cours de différents mois de l'étude. Ces périodes étaient définies à l'aide de renseignements sur le lever et le coucher du soleil ainsi que sur les habitudes d'utilisation prévues pour les divers types de sites. Les mois d'hiver (de novembre à avril) incluaient deux périodes par jour, alors que les mois d'été (de mai à octobre) en comprenaient trois.

La (sous) stratification temporelle de chaque échantillon mensuel a été effectuée selon les jours de fin de semaine par opposition aux jours de semaine. Il vaut la peine de faire remarquer que la strate des jours de fin de semaine incluait aussi les principaux jours fériés.

4.2 Sélection de l'échantillon

Comme on l'a signalé plus haut, nous avons choisi, au premier degré, un échantillon indépendant de paires site-période pour chacun des 12 mois de l'étude. Chaque échantillon mensuel comprenait deux composantes: a) un échantillon aléatoire stratifié de "n" sites et b) un échantillon aléatoire stratifié de "n" périodes.

Après la sélection des périodes pour chaque mois, les sites échantillonnés ont été attribués aléatoirement aux périodes sélectionnées. L'attribution des sites aux périodes a été faite entièrement au hasard pour les mois de février à juin, mais elle a été modifiée au cours des mois ultérieurs. À partir de juillet, un échantillon des périodes était choisi indépendamment pour chaque strate de type de site, l'attribution aléatoire se faisant dans la strate. La répartition des périodes échantillonnées (p. ex., le nombre de périodes en matinée et le nombre de périodes en soirée incluses dans

2) Pour les grappes de parcs dans la dernière de ces deux strates (printemps et automne), cinq périodes de deux mois qui ne se chevauchaient pas constituaient la base de sondage temporelle: janvier-février, mars-avril, mai-juin, septembre-octobre, novembre-décembre. Il faut remarquer que la période de deux mois de l'été, juillet-août (c.-à-d. le plein été), était exclue de la base de sondage pour ces parcs afin d'assurer le choix d'autres mois ne faisant pas partie de l'été. Pour chaque grappe dans cette strate, une parmi ces cinq périodes de deux mois était choisie avec PPT.

3.3 Troisième degré d'échantillonnage

Au troisième degré, les unités d'échantillonnage étaient les paires sortie-jour. Il est plus facile d'envisager l'échantillonnage au troisième degré comme la combinaison de deux choix indépendants: a) le choix de jours dans la période (bloc de deux mois) tirée au deuxième degré pour le parc (et b) le choix des sorties à partir d'une liste de sorties précises pour le parc. L'attribution aléatoire des jours échantillonnés aux sorties échantillonnées constituait le dernier degré.

L'échantillon des jours était stratifié par jours de semaine par opposition aux jours de fin de semaine (y compris les principaux jours fériés). Les jours échantillonnés étaient choisis avec probabilités égales dans chacune de ces deux strates. L'échantillon des sorties a été choisi avec probabilités proportionnelles à la taille (PPT). La mesure de taille attribuée à chaque sortie était l'utilisation relative de la sortie parmi toutes les sorties relevées pour le parc choisi. Cette mesure de l'utilisation a été obtenue avec l'aide d'employés de chacun des parcs.

Pour le plan d'échantillonnage du troisième degré, on faisait la distinction entre deux groupes de parcs désignés comme étant très achalandés (trois parcs: Grand Canyon et les deux parcs d'Hawaï) et peu achalandés (les 36 autres parcs de l'échantillon). La taille des échantillons pour les parcs peu achalandés était de 10 sorties et de 10 jours et, par conséquent, de 10 paires sortie-jour. Dans les trois parcs très achalandés, 15 sorties et 15 jours ont été choisis. Pour les parcs peu achalandés, nous avons utilisé une attribution égale entre la strate des jours de semaine et celle des jours de fin de semaine: 5 jours de semaine et 5 jours de fin de semaine ont été choisis de façon indépendante dans chacun de ces (36) parcs dans l'échantillon. Pour les parcs très achalandés, l'attribution était à peu près égale puisqu'on a choisi 7 jours de fin de semaine et 8 jours de semaine.

3.4 Quatrième degré d'échantillonnage

Au cours du quatrième degré, les visiteurs des parcs ont été interviewés sur le vif aux sorties choisies lors des jours sélectionnés. Un échantillon aléatoire systématique de visiteurs (ou de groupes de visiteurs) quittant le parc a été choisi avec un pas de sondage fixe pour chaque unité (paire sortie-jour) sélectionnée au troisième degré. On a laissé le pas de sondage varier d'un jour à l'autre afin de tirer profit de l'expérience des jours précédents et de la variabilité dans le nombre de visiteurs entre les jours et les sorties. Chaque visiteur faisant partie des groupes choisis admissibles était

soumis à un filtrage afin de s'assurer qu'il était admissible, puis interviewé s'il l'était (visiteur adulte).

3.5 Enquête sur l'utilisation des installations de l'arrière-pays et enquête postale

Le même échantillon de parcs (UPB) a été utilisé pour une enquête postale auprès des utilisateurs des installations de l'avant-pays et de celles de l'arrière-pays pendant la période de deux mois sélectionnée pour le parc. L'échantillon de l'enquête sur l'utilisation des installations de l'arrière-pays était limité au sous-ensemble des parcs échantillonnés pour lesquels il y avait une certaine utilisation de l'arrière-pays. Dans chacun de ces parcs échantillonnés, la base de sondage (du troisième degré) pour cette composante de l'enquête était fondée sur les permis de circulation dans l'arrière-pays délivrés pendant la période de la collecte des données (bloc de deux mois) établie pour le parc. Le dernier échantillon d'utilisateurs d'installations de l'arrière-pays était alors choisi avec probabilités égales à partir des listes de permis fournies par le personnel du parc. Un sous-échantillon de visiteurs choisis lors des interviews sur le vif était aussi sélectionné pour l'enquête postale.

L'échantillon utilisé pour l'enquête postale était fondé sur le même échantillon, pour le troisième degré, de paires sortie-jour choisies pour les interviews sur le vif. Pour chaque paire sortie-jour choisie, on a obtenu un sous-échantillon composé d'un nombre fixe de groupes parmi les groupes répondant aux interviews sur le vif (ce nombre était de 15 dans les parcs très achalandés et de 10 dans les parcs peu achalandés). Un autre degré de sous-échantillonnage fut de choisir une personne parmi les répondants dans chaque groupe sous-échantillonné. Ce sous-échantillonnage des groupes et des personnes s'effectuait avec échantillonnage aléatoire stratifié pour contrôler la composition démographique de l'échantillon final.

4. PLAN D'ÉCHANTILLONNAGE DE L'ENQUÊTE PORTANT SUR LES TROIS RIVIÈRES

4.1 Construction de la base de sondage et stratification

Tout d'abord, RTI et Terrestrial Environment Specialists (TES) ont construit une base de sondage fondée sur un relevé de tous les sites dans la région visée par l'enquête portant sur les trois rivières. Puis un échantillon stratifié à plusieurs degrés a été choisi indépendamment pour chacun des douze mois de l'étude. Les unités d'échantillonnage du premier degré étaient les paires site-période et celles du deuxième degré, les personnes s'adonnant à des activités récréatives dans les paires site-période choisies. Au premier degré de l'échantillonnage des périodes et des sites, on a eu recours à la stratification temporelle et à la stratification dans l'espace.

La stratification primaire dans la dimension temporelle a été effectuée selon le mois, et la stratification primaire des sites s'est faite selon le type d'utilisation: les sites (points d'accès) ont été classés comme sites de navigation

c) Dans la strate 3, “printemps et automne” (comprenant 4 grappes), la base de sondage constituée de blocs de deux mois renfermait 5 blocs de deux mois qui ne se chevauchaient pas.

Document 1

Grappes de parcs, strates du deuxième degré et blocs de deux mois sélectionnés

Strate 1: Plein été*		Période choisie
Stratum 2: Début et fin de l'été		
Grappe 2: Mount Rainier, N. Cascades, Olympic	Grappe 10: Sleeping Bear, Perry's Victory	Juillet-août
Grappe 4: Glacier, Yellowstone		Juillet-août
Grappe 10: Sleeping Bear, Perry's Victory		Juillet-août
Stratum 3: Printemps et automne		
Grappe 1: Haleakala, Hawaii Volcanoes	Grappe 7: Lake Mead, Saguaro, Casa Grande	Mars-avril
Grappe 9: Hot Springs, Wilson's Creek, Buffalo	Grappe 14: Cumberland Isd., Canaveral, Everglades, Gulf Island	Mars-avril
Grappe 12: Shenandoah, Frederickicksburg, Assateague		Juin-juillet
Grappe 13: Great Smoky, Cape Hatteras, Fort Sumter		Juin-juillet
Grappe 11: Cape Cod, Delaware Gap, Gettysburg		Août-septembre
Grappe 8: Bandelier, Lake Meredith		Mai-juin
Grappe 6: Glen Canyon, Grand Canyon, Walnut Canyon		Juin-juillet
Grappe 5: Dinosaur, Rocky Mtn., Mt. Rushmore		Juin-juillet
Grappe 3: Lassen, Yosemite, Kings Canyon/Sequoia		Août-septembre

*Choix de juillet-août avec probabilités égales à l'unité pour chaque grappe dans cette strate.

Deux raisons justifiaient cette stratification temporelle: elle permettait d'assurer que la collecte des données s'effectuait pendant toute l'année et de distinguer les parcs où la grande majorité des visites ont lieu pendant l'été de ceux où la répartition des visites est plus uniforme. Nous avons utilisé des procédures différentes, décrites ci-après, pour choisir les blocs de deux mois dans les deux dernières strates.

1) Pour les grappes de parcs dans la première de ces deux strates (début et fin de l'été), onze périodes qui se chevauchaient étaient incluses dans la base de sondage temporelle; janvier-février, février-mars, . . . , novembre-décembre. Une de ces périodes était alors choisie avec probabilités proportionnelles à la taille (PPT). Il faut remarquer que, pour cette strate, chaque mois était inclus dans deux périodes faisant partie de la base de sondage sauf janvier et décembre. La probabilité de sélection de ces deux mois d'hiver était donc encore réduite (d'avantage que la probabilité déjà faible attribuée aux périodes hivernales avec la procédure PPT).

de beaucoup inférieur à celui utilisé dans les autres strates. C'est dans la strate urbaine que l'on observe le taux d'échantillonnage le plus faible (1 sur 79).

La strate à tirage complet comprenait les sept parcs qui, à cause des prescriptions de la loi, devaient être inclus dans l'étude. De plus, elle comprenait les parcs dont les taux de visites globaux (annuels) étaient suffisamment élevés pour assurer leur tirage dans l'échantillon obtenu au premier degré. La liste des 39 parcs dans l'échantillon est présentée dans le document 1; un 40^e parc sélectionné (Grand Teton) a été éliminé de l'échantillon pour des raisons politiques.

Les calculs d'optimisation du plan ont mené, pour le premier degré, à une taille d'échantillon d'environ 40 parcs, donnant un total de 405 paires site-période (ou, dans le présent cas, de paires sortie-jour) choisies parmi les parcs très achalandés et peu achalandés. Comme on le décrit dans la sous-section 3.3, 15 paires sortie-jour ont été choisies dans chacun des trois parcs très achalandés et 10 dans chacun des 36 parcs peu achalandés dans l'échantillon. Pour une optimisation exacte, il faudrait utiliser les composantes de la variance pour la variance entre les parcs et pour la variance à l'intérieur des parcs. Ces composantes de la variance ont été obtenues de façon approximative à partir de données tirées d'une enquête antérieure et des données mensuelles sur les visites, qui sont disponibles pour toutes les zones administrées par le NPS.

3.2 Choix des blocs de deux mois

Bien que le choix des blocs de deux mois ait été effectué avec probabilités proportionnelles à la taille (PPT), on a aussi tenu compte d'exigences pratiques.

Premièrement, il était souhaitable que la formation des employés dans chacun des parcs s'effectue immédiatement avant la collecte des données. On a constitué des grappes géographiques de parcs afin que les moniteurs puissent visiter tous les parcs d'une grappe au cours d'un seul voyage. Plus précisément, les 39 parcs dans l'échantillon ont été groupés en 14 grappes de ce genre. Cette exigence a mené à la sélection d'une période de deux mois pour chaque grappe de parcs. Ainsi, quatorze blocs de deux mois ont été choisis, un pour chaque grappe. La mesure globale de visites dans tous les parcs de la grappe.

Le document 1 montre les parcs échantillonnés dans chaque grappe ainsi que le bloc de deux mois choisi pour la grappe (c.-à-d. pour chaque parc dans la grappe). Trois strates, de groupes de grappes et par conséquent de parcs, ont été formées pour le choix des blocs de deux mois:

a) Dans la strate 1, “plein été” (comprenant 3 grappes), la base de sondage constituée de blocs de deux mois ne renfermait que la période de pointe de l'été, soit juillet et août, qui était alors choisie avec probabilités égales à un pour ces grappes;

b) Dans la strate 2, “début et fin de l'été” (composée de 7 grappes), la base de sondage constituée de blocs de deux mois renfermait 11 blocs de deux mois qui se chevauchaient;

chaque mois et pour chaque strate (p. ex., selon le genre: navigation de plaisance, pêche sportive ou fréquentation des parcs). On permettait aux bases de sondage spatiales et temporelles de varier d'un mois à l'autre. On traite de la stratification ainsi que de la répartition pour cet échantillon dans la section 4.

Par opposition à la situation qui vient d'être décrite, dans la base de sondage temporelle de l'enquête auprès des visiteurs du NPS, on considérait tout d'abord des blocs de deux mois pour chaque parc dans l'échantillon (UPB).

L'utilisation de périodes de deux mois comme unités d'échantillonnage (au deuxième degré) dans le temps permettait d'atteindre de façon efficace les objectifs de l'enquête pour deux raisons fondamentales. Premièrement, elle permettait une concentration (géographique) efficace des ressources en personnel et répartissait la collecte des données sur toute l'année. Deuxièmement, ce choix de périodes permettait de saisir les fluctuations saisonnières des visites dans tout le réseau des parcs, fluctuations qui découlent du fait que, dans certains parcs, le nombre de visiteurs est relativement plus élevé au printemps alors que pour d'autres, il est plus élevé en automne et ainsi de suite.

Un bloc de deux mois a été choisi pour chaque parc dans l'échantillon afin que la collecte des données puisse être concentrée dans le temps de façon efficace. Puis, à l'étape suivante de la sélection dans le temps, des jours ont été choisis dans le bloc de deux mois pour chaque parc dans l'échantillon. Comme les parcs eux-mêmes, au premier degré, ces blocs de deux mois ont été choisis avec probabilités proportionnelles à la taille (PPT), la mesure de taille étant l'ensemble des visiteurs.

On a équilibré avec soin la taille de l'échantillon ainsi que sa répartition entre les différents degrés d'échantillonnage afin de minimiser les effets de groupement associés aux grappes dans le temps et dans l'espace. Pour l'étude portant sur les trois rivières, ce groupement se produit au premier degré de la sélection, quand les unités d'échantillonnage sont les sites et les périodes. Dans la répartition, on a aussi tenu compte du fait que la taille des échantillons varie dans des échantillons mensuels indépendants successifs. Pour l'enquête du NPS, les grappes temporelles provenaient des blocs de deux mois et des jours choisis aux différents degrés. Les grappes géographiques provenaient de l'utilisation des parcs et des sorties de parcs comme unités d'échantillonnage pour cette enquête.

Dans la section ci-après, on décrit plus en détail le plan utilisé pour l'enquête du NPS.

3. PLAN D'ÉCHANTILLONNAGE DE L'ENQUÊTE DU NPS

Le plan d'échantillonnage de l'enquête auprès des visiteurs du NPS tirait profit d'informations supplémentaires de divers genres et provenant de différentes sources: - Renseignements obtenus lors d'études antérieures (p. ex., le classement des parcs fondé sur l'exposition au bruit ainsi que des classifications faites par des employés du NPS).

Dans les sous-sections ci-après, on décrit les divers degrés de la sélection pour les interviews sur le vif auprès des visiteurs. Cette partie de l'enquête sera aussi désignée comme l'enquête sur l'avant-pays pour la distinguer d'une enquête sur les utilisateurs des installations de l'arrière-pays, qui a été réalisée en même temps dans les parcs échantillonnés. Dans la sous-section 3.5, on décrit l'enquête sur l'utilisation des installations de l'arrière-pays ainsi qu'une enquête postale réalisée auprès d'un sous-échantillon de répondants à l'enquête sur l'avant-pays ainsi qu'auprès d'un échantillon d'utilisateurs des installations de l'arrière-pays. Le quatrième degré de la sélection, pour l'enquête postale, comprenait des degrés (et des phases) additionnels de sélection. Pour l'enquête comportant des interviews sur le vif auprès des visiteurs, des groupes (c.-à-d. des véhicules) étaient choisis comme grappe ultime: on demandait à toutes les personnes dans un groupe dans l'échantillon si elles acceptaient d'être interviewées. Pour l'enquête postale, on a tiré un sous-échantillon dans les groupes et une personne a été choisie dans chaque groupe faisant partie du sous-échantillon.

3.1 Construction de la base de sondage et premier degré d'échantillonnage

Nous avons construit une base de sondage pour le choix des parcs en compilant des renseignements du NPS sur les visites (mensuelles et annuelles) dans les parcs et sur l'exposition au bruit, renseignements qui étaient utilisés de deux façons distinctes dans le plan d'échantillonnage: pour la stratification et pour l'attribution des mesures de taille. Ces derniers renseignements étaient basés sur deux sources différentes: a) une étude antérieure du NPS dans laquelle les parcs étaient classés en fonction de l'exposition potentielle et b) une classification des parcs effectuée indépendamment par des employés du NPS (surintendants de parc, employés travaillant dans les régions, etc.). En consultation avec le NPS, RTI a combiné ces deux classifications pour l'exposition au bruit afin de construire onze strates. La stratification répartit les parcs en catégories très élevée, élevée, faible et très faible. Les strates ont été divisées en deux sous-strates à l'aide du classement dans la strate. (Il est à noter que la strate "moyenne" n'a pas été subdivisée à cause du faible nombre de parcs qu'elle contenait.)

En plus de l'exposition au bruit, ces strates incorporaient trois classes de parcs qui méritent un traitement distinct: 1) parcs urbains et de banlieue, 2) parcs pour lesquels des données sur les visites manquent (données nécessaires pour la sélection avec PPT) et 3) parcs dont la forme allongée présente des problèmes particuliers d'accès et réduit la signification des évaluations des expositions au bruit antérieures. Dans ces classes, le taux d'échantillonnage était

Ces deux enquêtes illustrent des questions telles que la façon dont les visiteurs choisissent au dernier degré de leur évaluation de visiter un site-période pour l'avant-dernier degré (en n'oubliant pas l'impact du site-période sur le premier degré et les autres degrés) et les raisons pour lesquelles les visiteurs choisissent de visiter un site-période pour le premier degré et les autres degrés. Les résultats de ces deux enquêtes sont présentés à la section 6.

2. APERÇU DES PLANS D'ÉCHANTILLONNAGE: POINTS COMMUNS ET DIFFÉRENCES

Pour les deux enquêtes, les derniers échantillons de visiteurs étaient choisis au moyen d'interviews sur le vif au moment où les visiteurs quittaient les lieux choisis dans l'échantillon pendant les périodes sélectionnées. Il fallait effectuer des interviews à la sortie afin que l'on puisse relever l'attitude des visiteurs immédiatement après qu'ils aient utilisé les installations de loisirs. De plus, dans les deux études, les visiteurs étaient choisis à partir d'échantillons de paires site-période. L'utilisation de paires site-période comme unités d'échantillonnage remonte à l'étude de Kish, Lovejoy et Rackow (1961). Cette définition des unités d'échantillonnage permet de choisir les visiteurs selon un calendrier de collecte des données qui précise quels sites seront vus à quel moment. Contrairement à l'enquête portant sur les trois rivières, le choix des paires site-période ne constituait pas le premier degré pour l'enquête du NPS. Les unités primaires d'échantillonnage (UPÉ) pour cette enquête étaient les zones du NPS, ou les parcs. L'enquête du NPS comprenait plusieurs degrés de sélection, qui sont décrits dans la prochaine section.

Jours dans chaque enquête étaient, eux, fort différents. Par exemple, pour l'étude portant sur les trois rivières, l'échantillon choisi était constitué de douze échantillons mensuels indépendants. Chaque échantillon mensuel avait essentiellement le même plan d'échantillonnage aléatoire stratifié, mais on utilisait une répartition différente de l'échantillon ainsi que des tailles d'échantillon différentes au cours des différents mois. Ce plan tenait compte des variations saisonnières dans les caractéristiques des loisirs et des sports et permettait le calcul d'estimations pour

comme unités primaires d'échantillonnage (UPB) à partir desquelles les visiteurs sont choisis au deuxième degré. Les exemples comprennent les sondages faits à la sortie de l'isoloir (voir, par exemple, Levy 1983), les interviews sur le vif dans les centres commerciaux (voir, par exemple, Sudman 1980) et d'autres enquêtes sur les transports et la circulation (Gough et Ghauri 1977; Kish, Lovejoy et Rakow 1961). Parmi les questions relatives au plan d'échantillonnage qui sont importantes dans les enquêtes auprès des visiteurs, on peut distinguer les problèmes généraux suivants:

- Il est souhaitable de choisir avec une plus grande probabilité les paires site-période pour lesquelles il y a un plus grand nombre de visiteurs, la stratification et le choix avec PPT sont alors des caractéristiques efficaces du plan.
- Les paramètres relatifs à la collecte des données sont la clé pour préciser la longueur de la période ainsi que les taux d'échantillonnage pour chaque paire site-période; p. ex., on doit faire certains compromis entre la possibilité que les employés sur le terrain soient trop occupés (périodes brèves, taux d'échantillonnage élevés) ou pas assez occupés (longues périodes, faibles taux d'échantillonnage).
- Les objectifs sur le plan de l'analyse ainsi que l'efficacité nous incitent à utiliser, pour la stratification temporelle, des dimensions telles que la saison, le mois, la fin de semaine par opposition au jour de semaine, et même à la période de la journée; p. ex., le besoin d'estimations saisonnières laisse supposer l'utilisation de saisons ou de mois comme strates pour le choix des périodes.

Les deux études dont on traite dans cet article ont un objectif principal commun, soit celui de caractériser la population des visiteurs dans la région ou dans le pays sur une année complète. Toutefois, dans chaque cas, elles diffèrent au niveau des estimations les plus importantes qui mènent aux caractéristiques fondamentales du plan. Dans l'étude sur les trois rivières, mais non dans l'étude du NPS, on devait calculer des estimations mensuelles assez précises. Pour la première étude, alors, les unités d'échantillonnage dans le temps, en l'occurrence les jours, ont été stratifiées en mois. La stratification dans l'espace des sites de l'étude sur les trois rivières était géographique et par type d'activités de loisirs (navigation de plaisance, pêche sportive ou fréquentation des parcs).

Pour l'étude du N°3, la stratification primaire a été faite selon le type de parc. Certains parcs devaient absolument être inclus dans l'échantillon pour répondre à des prescriptions de la loi. Dans ces parcs et dans d'autres parcs choisis, on souhaitait aussi calculer des estimations propres à chaque parc. Pour ces parcs, dits parcs très achalandés, on a alors décidé de choisir relativement plus de parcs site-période. Le problème d'optimisation pour le plan d'échantillonnage initial était de savoir comment répartir la taille de l'échantillon entre les degrés d'échantillonnage, c.-à-d. de déterminer combien de parcs et combien de paires site-période par parc devaient être choisis. Dans la section 3, on étudie une solution à ce problème qui est une fonction des corrélations à l'intérieur d'une grappe dans les parcs et dans les périodes. L'optimisation du plan

Enquêtes par sondage auprès des visiteurs

RONALDO IACHAN et SUZANNE S. KEMP¹

RÉSUMÉ

Dans le présent article, on traite de la conception d'enquêtes auprès des visiteurs. Pour illustrer le sujet, on décrit deux enquêtes récentes. La première est une enquête auprès des visiteurs des zones administrées par le National Park Service à l'échelle des États-Unis pendant toute une année (1992). La deuxième est une enquête menée auprès des personnes qui ont utilisé des installations de loisirs du bassin des trois rivières autour de Pittsburgh, en Pennsylvanie, pendant une période de douze mois. Les deux enquêtes comportaient un échantillonnage dans le temps avec stratification temporelle et stratification dans l'espace. Les unités d'échantillonnage avaient la forme de paires site-période à l'avant dernier degré d'échantillonnage soit celui précédant de l'échantillonnage des visiteurs. L'attribution aléatoire des sites de l'échantillon à des périodes permet d'obtenir des estimations non biaisées pour les strates temporelles (p. ex., des estimations mensuelles et saisonnières) ainsi que des estimations pour les strates définies selon la région et le genre d'utilisation.

MOTS CLÉS: Personne qui utilise des installations de loisirs; échantillonnage dans le temps; site-période.

Les enquêtes auprès des visiteurs présentent des défis particuliers dont on traite rarement dans les publications statistiques. Le présent article tente de combler cette lacune en décrivant la conception de deux enquêtes réalisées récemment par le Research Triangle Institute (RTI) et en mettant l'accent sur les caractéristiques communes de ces deux enquêtes. Nous espérons que les leçons apprises au cours de ces travaux profiteront aux chercheurs qui planifient des enquêtes semblables.

La première enquête était une étude des visiteurs des zones administrées par le National Park Service (NPS), réalisée conjointement pour le National Park Service par RTI et HBRS, Inc. Cette étude comportait la production d'un échantillon probabiliste de visiteurs des parcs, qui représentait les visiteurs de 323 zones administrées par le NPS dans tous les États-Unis (à l'exception de l'Alaska) pendant toute l'année 1992. Pour simplifier la rédaction, nous désignerons les zones administrées par le NPS par l'expression "parcs", tout en signalant que ces zones comprennent aussi des sites historiques ainsi que des parcs à vocation culturelle. Le principal objectif de l'enquête du NPS était d'évaluer les expériences et les problèmes des visiteurs en portant une attention particulière aux problèmes liés au survol des parcs par des avions (p. ex., le bruit et autres gênes possibles). Diverses données ont aussi été recueillies au moyen d'une enquête postale portant sur un sous-échantillon des visiteurs sélectionnés.

La deuxième enquête était une étude portant sur les personnes qui utilisent des installations de loisirs du bassin des trois rivières, dans la région de Pittsburgh, le long des rivières Monongahela, Allegheny et Ohio en 1992 (ou, plus précisément, entre février 1992 et janvier 1993). Cette enquête a été réalisée conjointement pour la Ohio River

Valley Sanitation Commission par RTI et Terrestrial Environmental Specialists. La région visée par l'étude comprenait un segment de 40 milles de la rivière Ohio, un segment de 24 milles de la rivière Monongahela ainsi qu'un segment de 7 milles de la rivière Allegheny. L'objectif principal de l'étude sur les trois rivières était de construire un profil de base des utilisateurs des installations de loisirs de la région et de modéliser la valeur économique que ces derniers attribuent à diverses activités. On a fait la distinction entre trois types de base d'activités de loisirs et de sports: navigation de plaisance, pêche sportive et fréquentation des parcs.

L'étude portant sur les trois rivières est la plus complète d'une série d'études réalisées par RTI pour évaluer l'impact sur l'environnement dans un certain nombre d'États. Ces études estiment la réduction possible de la valeur économique ou récréative attribuée par les personnes qui utilisent effectivement ou qui pourraient utiliser des installations de loisirs à des zones qui ont été ou qui pourraient être touchées. Bien que, pour une enquête plus étendue portant sur les utilisateurs éventuels de telles zones, on puisse considérer un plan d'échantillonnage dans le cadre duquel les répondants sont interviewés par téléphone, on a trouvé qu'il fallait avoir recours à des interviews sur le vif auprès des visiteurs afin d'obtenir de ces derniers des renseignements à un moment très rapproché de celui où ils utilisent effectivement les installations.

Une discussion des problèmes liés à la conception d'enquêtes auprès des visiteurs comme celles qui viennent d'être mentionnées a été présentée récemment dans Kalton (1991), notamment en ce qui a trait à l'échantillonnage dans le temps et dans l'espace dont l'importance est critique dans notre cadre de travail. Sous sa forme la plus simple, un plan d'échantillonnage prototype à deux degrés pour une enquête auprès des visiteurs considère des paires site-période

¹ Ronaldo Iachan, Statisticien chercheur et Suzanne S. Kemp, Statisticienne, Research Triangle Institute, 3040 Cornwallis Road, Research Triangle Park, NC 27709-2194, U.S.A.

- GWET, J.-P., et RIVEST, L.-P. (1992). Outlier resistant alternatives to the ratio estimator. *Journal of the American Statistical Association*, 87, 1174-1182.
- HAMPPEL, F.R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69, 383-393.
- HAMPPEL, F.R., RONCHETTI, E.M., ROUSSSEUW, P.J., et STAHEL, W.A. (1986). *Robust Statistics*. New York: Wiley.
- HANSEN, M.H., MADOW, W.G., et TEPPING, B.J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association*, 78, 776-807.
- HIDROGLOU, M.A., et SRINATH, K.P. (1981). Some estimators of a population total from simple random samples containing large units. *Journal of the American Statistical Association*, 76, 690-695.
- HUBER, P.J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35, 73-101.
- HULLIGER, B. (1991). Nonparametric M-estimation of a population mean. Thèse de doctorat ETH No. 9443, ETH Zürich.
- IACHAN, R. (1984). Sampling strategies, robustness and efficiency: the state of the art. *Revue Internationale de Statistique*, 52, 209-218.
- JAECKEL, L.A. (1971). Robust estimates of location: symmetry and asymmetric contamination. *Annals of Mathematical Statistics*, 42, 1020-1034.
- KISH, L. (1965). *Survey Sampling*. New York: Wiley.
- LITTLE, R.J.A., et SMITH, Ph.J. (1987). Editing and imputation for quantitative survey data. *Journal of the American Statistical Association*, 82, 58-68.
- OEHLERT, G.W. (1985). The random average mode estimator. *Annals of Statistics*, 13, 1418-1431.
- RAO, J.N.K. (1966). Alternative estimators in PPS sampling for multiple characteristics. *Sankhya A*, 28, 47-60.
- RIVEST, L.-P. (1993). Winsorization of survey data. *Actes de la 49ème Session, Institut International de Statistique*.
- SEARLS, D.T. (1966). An estimator for a population mean which reduces the effect of large observations. *Journal of the American Statistical Association*, 61, 1200-1204.
- SHOEMAKER, L.H., et ROSENBERGER, J.L. (1983). Moments and efficiency of the median and trimmed mean for finite populations. *Communications in Statistics, Simulations and Computations*, 12(4), 411-422.
- SMITH, T.M.F. (1987). Influential observations in survey sampling. *Journal of Applied Statistics*, 14, 143-152.
- STATISTICAL SCIENCES, INC. (1990). *S-PLUS Software*. Seattle: Statistical Science, Inc.

REMERCIEMENTS

On peut conclure de cette simulation limitée que l'estimateur REM perd peu en terme d'EQM comparativement à l'estimateur HT lorsqu'il n'y a pas de valeurs aberrantes dans la population. Il gagne modérément en efficacité dans les populations contenant des valeurs aberrantes symétriques, et beaucoup plus lorsque les valeurs aberrantes sont asymétriques. L'estimateur HTR perd plus d'efficacité que l'estimateur REM dans les situations idéales. L'adaptabilité des estimateurs REM est profitable.

Hulliger (1991) confirme ces conclusions à l'aide de vastes simulations utilisant des populations infinies. Il montre que les gains autorisés par les estimateurs robustes peuvent être considérables dans les cas de populations asymétriques comportant des valeurs aberrantes. Toutefois, les gains possibles d'efficacité envisageables avec les estimateurs robustes disparaissent avec les gros échantillons puisque le biais domine l'EQM dans de tels cas. Par contre, si les valeurs aberrantes qui apparaissent dans un échantillon donné ne sont pas représentatives (p. ex., s'il s'agit d'erreurs de codage non corrigées), les estimateurs robustes seront beaucoup plus efficaces que l'estimateur HT pour toutes les tailles d'échantillons.

BIBLIOGRAPHIE

- CHAMBERS, R.L. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association*, 81, 1063-1069.
- COCHRAN, W.G. (1977). *Sampling Techniques* (3ème Ed.). New York: Wiley.
- DEY, A., et SRIVASTAVA, A.K. (1987). Méthode d'échantillonnage avec probabilités de sélection proportionnelles à la taille. *Techniques d'enquête*, 13, 93-100.
- FULLER, W.A. (1991). Simple estimators of the mean of skewed populations. *Statistica Sinica*, 1, 137-158.
- GLASSER, G.J. (1962). On the complete coverage of large units in a statistical study. *Revue de l'Institut International de Statistique*, 30, 28-32.
- GODAMBE, V.P. (1955). A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society, Series B*, 17, 269-278.

Pour chacune des populations, un ensemble de 400 échantillons a été tiré pour évaluer les estimateurs. La précision obtenue est suffisante pour tirer des conclusions (voir les écarts-types des valeurs de l'efficacité, dans le tableau 1).

Tableau 1
Simulations de Monte-Carlo avec estimateurs HTR et REM

Populations									
GODA	HTLS	HTE	HTG	HMT	HMTE				
Moy. M.-C. de HT	6.996	4.531	4.483	2.271	1.068	0.991			
Biais rel. de HTR	-0.002	-0.001	-0.009	-0.009	0.006	-0.052			
Biais rel. de REM	0.000	-0.001	-0.007	-0.008	-0.002	-0.035			
E.-t. rel. de HT	0.067	0.041	0.044	0.098	0.107	0.170			
E.-t. rel. de HTR	0.070	0.044	0.040	0.087	0.117	0.144			
E.-t. rel. de REM	0.068	0.042	0.040	0.091	0.107	0.146			
Eff. de HTR	0.911	0.876	1.110	1.310	0.827	1.234			
Eff. de REM	0.969	0.981	1.158	1.194	0.989	1.284			
E.-t. M.-C. de l'eff. HTR	0.020	0.017	0.073	0.009	0.018	0.001			
E.-t. M.-C. de l'eff. REM	0.003	0.009	0.037	0.002	0.013	0.002			

NOTA: Le biais relatif (Biais rel.) et l'écart-type relatif (E.-t. rel.) correspondent aux valeurs du biais et de l'écart-type divisées par la moyenne M.-C. de l'estimateur HT. Les valeurs de l'efficacité (Eff.) correspondent à l'EQM de l'estimateur HT divisée par l'EQM de l'estimateur. Les valeurs estimées de l'écart-type de ces estimations de Monte-Carlo sont indiquées dans les deux dernières lignes.

Les résultats sont présentés dans le tableau 1. La distorsion relative de l'estimateur HTR est toujours plus grande que celle de l'estimateur REM. Les distorsions des estimateurs possèdent le même signe, sauf lorsqu'elles sont très petites. Exception faite des populations HTE et HMTE, la variance de l'estimateur HTR est plus grande que celle de l'estimateur REM. Si l'estimateur HTR perd 9% de son efficacité avec la population GODA, alors que l'estimateur HT devrait être optimal, l'estimateur REM en perd très peu. Avec la population HTLS, où l'estimateur HT est l'estimateur des moindres carrés, l'estimateur HTR perd environ 12% d'efficacité. Ici encore, la perte d'efficacité de l'estimateur REM est limitée. La population HTG contient les valeurs aberrantes résiduelles asymétriques. L'estimateur HTR gagne environ 11% d'efficacité (noter cependant l'erreur de 7.3%) et l'estimateur REM en gagne environ 16%. Avec les valeurs aberrantes asymétriques de la population HTE, le gain de l'estimateur HTR est d'environ 31%, tandis que celui de l'estimateur REM est de 19%. Dans un cas où ni la régression par l'origine, ni la symétrie des erreurs ou la proportionnalité de leurs variances par rapport à la variable explicative n'ont d'empêchement (c.-à-d., pour la population HMT), l'estimateur HTR perd 17% d'efficacité comparativement à l'estimateur HT tandis que l'estimateur REM conserve pratiquement toute son efficacité. Si dans une telle population on vient à relever quelques valeurs aberrantes asymétriques comme dans la population HMTE, les deux estimateurs robustes gagneront énormément d'efficacité par rapport à l'estimateur HT: 23% et 28% respectivement.

gamma dont la variance est proportionnelle à $x^{3/2}$. Ainsi, la variable y présente la distribution proposée par Hansen, Madow et Teping (1983, p. 781). Finalement, une population $y^{(6)}$ (HMTE) est générée avec 120 observations tirées de la même distribution que $y^{(5)}$, mais comportant 8 observations choisies au hasard de la distribution $\text{Exp}(2)$. Les six populations décrites ci-haut sont choisies pour leur réalisme. Elles utilisent toutes la même population de valeurs x (voir figure 1).

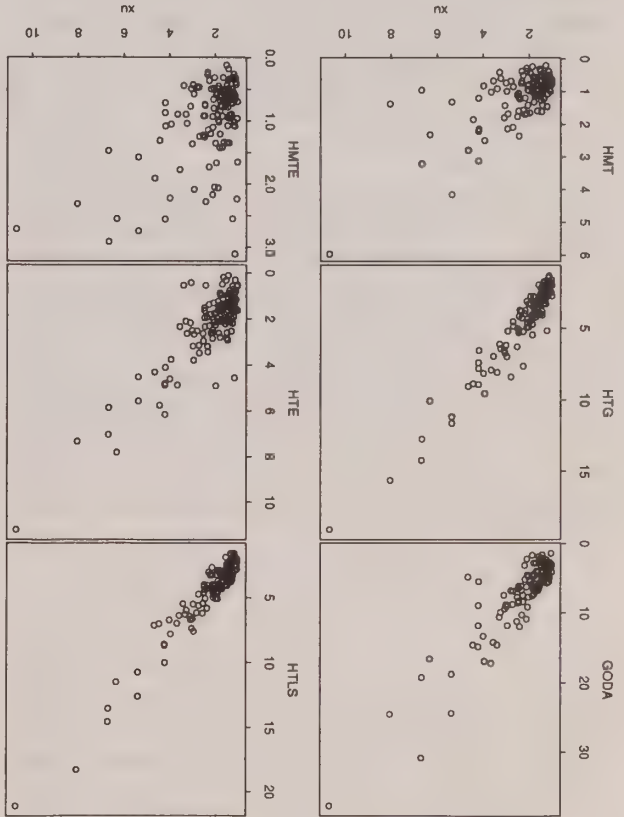


Figure 1. Populations de l'étude Monte-Carlo.

L'estimateur HTR de la simulation utilise

$$\eta(x_i', r_i') = w(x_i', k_x) \psi_{\text{HUB}}(r_i', k_r \text{ med}_S | r_i' |),$$

avec $w(x_i', k_x) = \min(1, k_x \text{ med}_U | x_i' | / | x_i' |)$ et $k_x = 2$. La fonction de pondération $w(x_i', k_x)$ correspond à une fonction Huber asymétrique $\psi_{\text{HUB}} = \min(x_i', k_x)$ qui ne donne un poids moins grand qu'aux grandes valeurs absolues évaluées dans la solution de l'itération précédente de l'algorithme par les moindres carrés répondérés par itérations. L'estimateur REM utilise le même η avec les constantes de mise au point k_x, k_r , évaluées à 20 points, qui se trouvent sur la diagonale de l'étendue de k_x et de k_r . On peut s'adresser à l'auteur pour obtenir les fonctions S-PLUS utilisées pour le calcul des estimateurs.

La troisième proposition de Huber (Huber 1964, p. 97) et la moyenne tendancielle adaptée de Jaekel (Jaekel 1971) s'adressent à des variables aléatoires symétriques et utilisent donc une estimation de la variance au lieu d'une estimation de l'EQM. Les estimateurs REM sont semblables, mais leur objectif est d'estimer la moyenne des distributions asymétriques.

Nous abordons ici l'utilisation des estimateurs REM pour les plans d'échantillonnage PIPT. Considérons un ensemble paramétrique de fonctions $\{\eta_k(x, r) : k \in K\}$, où $K \subset \mathbf{R}^p$ représente l'ensemble de paramètres. Habi-tuellement, $p = 1$ ou 2 pour rendre la minimisation faisable et pour limiter la perte d'efficacité due à l'estimation du paramètre de nuisance k . Nous n'appellerons pas k un paramètre, mais plutôt une constante de mise au point, pour éviter toute confusion avec la notion de paramètres utilisée dans le cadre des distributions de probabilité. Un ensemble propre de fonctions η nous donne un ensemble $\mathcal{B} := \{\beta(F_S, \eta_k) : k \in K\}$, où $\beta(F_S, \eta_k)$ est la pente d'un estimateur HTR. Pour assurer la cohérence de l'estimateur REM, nous supposons que $\lim_{k \rightarrow \infty} \eta_k(x, r) = r \vee (x, r)$ de sorte que l'estimateur HT est un élément de \mathcal{B} . L'EQM de $\beta(F_S, \eta_k)$ peut être estimée par

$$r(F_S, k) = \max(v_r(F_S, k), 0) + (\beta(F_S, k) - \beta_{MC}(F_S))^2, \quad (12)$$

où $v_r(F_S, k)$ est l'estimateur de variance (9) ou un autre estimateur de la variance de $\beta(F_S, \eta_k)$. Nous utilisons $\max(v_r, 0)$ dans $r(F_S, k)$ parce que l'estimateur de variance (9) pourrait devenir négatif. Typiquement, la fonction $r(F_S, k)$ avec $k \in \mathbf{R}_+$ possède un maximum égal ou presque égal à $k = 0$, à cause d'une distorsion importante. Elle devient ensuite minimale lorsque le biais et la variance sont toutes les deux petites. Pour les constantes de mise au point grandes, $r(F_S, k)$ s'approche de la variance de l'estimateur HT, habituellement par le bas.

Définition 3. Supposons que $r(F_S, \cdot)$ présente un minimum global à $k^m(F_S) \in K$. Dans ce cas, l'estimateur REM de la moyenne de la population est $M(F_S) = x_U \beta(F_S, \eta_{k^m})$. Les estimateurs REM à fonctions de définition adéquates présentent une équivalence d'échelle et n'ont pas besoin d'un estimateur d'échelle. Il s'agit en général d'estimateurs cohérents de la moyenne de la population. Hülliger (1991, chapitre 2) fournit la preuve de la grande cohérence des estimateurs REM de l'espérance d'une variable aléatoire.

Les problèmes posés par le caractère non unique du minimum ou par les cas où le minimum n'est pas atteint pour K sont facilement résolus en pratique par l'inspection de la fonction $r(F_S, k)$. (S'il existe plusieurs minimums globaux, choisir celui assorti de la plus petite constante de mise au point pour obtenir une plus grande robustesse.) La partie biaisée de $r(F_S, k)$ implique la pente $\beta_{MC}(F_S)$ de l'estimateur HT. Cet élément assure le transfert de la sensibilité de l'estimateur HT aux estimateurs REM, et la robustesse des estimateurs HTR est encore une fois perdue. Par contre, si l'estimateur REM doit être cohérent pour la moyenne de la population, il n'y a aucun moyen d'éviter un estimateur cohérent et, par conséquent, non robuste,

5. BRÈVE ÉTUDE DE SIMULATION

Nous présentons dans cette section des simulations réalisées avec des populations de taille $N = 128$ et des échantillons de taille $n = 16$. Nous utilisons le plan d'échantillonnage proposé par Dey et Srivastava (1987) (à noter qu'il manque un facteur de 2 dans leur formule (2.3)). Dey et Srivastava proposent de constituer des groupes $m > n/2$. Les totaux de groupes $\sum_{j=1}^m x_{ij}$ ($i = 1, \dots, m$) doivent répondre à l'inégalité $\sum_{j=1}^m x_{ij} / \sum_{j=1}^m x_i > (n - 2) / (n(m - 1))$. Ainsi, on ne tolère pour ces groupes qu'une très faible variabilité, et ils sont difficiles à former, en particulier dans le cas des plus grands échantillons (Hülliger 1991, p. 179).

Les x_i ($i = 1, \dots, N$) sont des réalisations indépendantes selon une distribution exponentielle à contamination d'échelle de 50% ayant son origine à 1, c.-à-d., $(X_i - 1) \sim 0.95 \text{Exp}(1) + 0.05 \text{Exp}(3)$, où $\text{Exp}(\theta)$ désigne la fonction de distribution exponentielle $1 - \exp(-x/\theta)$. Le facteur $+1$ est introduit pour abaisser la probabilité de réponses négatives dans la régression par l'intermédiaire du modèle d'origine à erreurs symétriques.

La première réponse, $y_i^{(1)}$, désignée par l'acronyme GODA, est une réalisation des variables normales indépendantes distribuées selon la formule $Y_i \sim \mathcal{N}(100x_i, x_i^2)$. Il s'agit du modèle en vertu duquel l'estimateur HT est optimal (voir Godambe 1955). La réponse $y_i^{(2)}$ (HTLS) est une réalisation des variables indépendantes distribuées selon la formule $Y_i \sim \mathcal{N}(2x_i, x_i/4)$. Il s'agit du modèle idéal où l'estimateur HT est un estimateur MC. Une troisième réponse, $y_i^{(3)}$ (HTG) est créée par le modèle $Y_i \sim 0.95\mathcal{N}(2x_i, x_i/4) + 0.05\mathcal{N}(2x_i, 9x_i/4)$. Les valeurs aberrantes résiduelles ont une échelle trois fois plus grande. La réponse $y_i^{(4)}$ (HTE) comporte des valeurs aberrantes asymétriques qui ne sont pas liées à la variable x . Les données (120 observations) proviennent en majorité de la distribution $Y_i \sim \mathcal{N}(2x_i, x_i/4)$ de $y_i^{(2)}$, mais 8 observations choisies au hasard viennent de la distribution $\text{Exp}(2.5)$. La population $y_i^{(5)}$ (HMT) vient d'une distribution assortie d'une espérance $0.4 + 0.25x_i$; elle a une distribution

Il s'agit là d'un problème général de l'estimation robuste pour les sous-populations (domaines) puisque la définition d'une valeur aberrante dépend de la population de référence. Une observation peut constituer une valeur aberrante dans une sous-population particulière tout en étant parfaitement anodine dans une autre. Ainsi, un estimateur robuste convenant à une sous-population donnée peut très bien donner de mauvais résultats avec une autre. Souvent, aucun accroissement de la robustesse n'est nécessaire ni souhaitable pour les moyennes générales alors qu'il s'avère essentiel pour les moyennes des sous-populations par suite de l'apparition de valeurs aberrantes. Heureusement, la taille de l'échantillon est souvent beaucoup moindre dans les sous-populations que dans les populations entières, et la composante du biais de l'erreur quadratique moyenne d'un estimateur robuste est relativement plus petite que la composante de la variance. Ainsi, les estimateurs robustes sont peut-être plus efficaces que les estimateurs HT lorsqu'on les utilise pour l'estimation de domaines.

3.2 Stratégie de Hansen-Hurwitz

Lorsqu'on utilise un échantillonnage avec remise et à probabilités de sélection inégales, on préfère l'estimateur Hansen-Hurwitz à l'estimateur HT. On peut accroître la robustesse de l'estimateur Hansen-Hurwitz d'une manière analogue à celle employée pour l'estimateur HT (voir Hulliger 1991, section 4.4) puisque le modèle sous-jacent est le même. L'approximation de la variance pour l'estimateur HH à robustesse accrue est plus simple que dans le cas de l'estimateur HTR car les produits croisés s'annulent par suite de l'échantillonnage avec remise du plan Hansen-Hurwitz.

3.3 Plan PIPT à robustesse accrue

Les rapports y_i/π_i de l'estimateur HT se comportent comme les termes de la somme d'une moyenne arithmétique. Une situation où les valeurs de π_i sont faibles et les valeurs de y_i élevées aura tendance à exagérer l'estimateur HT. Pour accroître la robustesse du plan d'échantillonnage contre des probabilités d'inclusion très grandes ou très petites, on peut utiliser $\hat{\pi}_i = nx_i/\sum U x_i$, où $x_i = \bar{x}_U + \psi^{\text{Hub}}(x_i - \bar{x}_U, k)$. Ainsi, la variable auxiliaire x_i est "Huberisée" de sa moyenne afin d'éviter les valeurs trop élevées ou trop faibles. Tirons maintenant un échantillon PIPT assorti de probabilités d'inclusion $\hat{\pi}_i$. L'estimateur HT est toujours $T^{\text{HT}} = (1/N) \sum y_i/\hat{\pi}_i$ et n'est toujours pas biaisé. Il n'est évidemment pas à l'épreuve des valeurs aberrantes en y et risque de perdre son efficacité si l'espérance du y_i n'est pas proportionnelle à $\hat{\pi}_i$. L'estimateur MC pondéré du modèle de superpopulation pour l'estimateur HT (voir section 2.1) avec probabilités d'inclusion $\hat{\pi}_i$ et une variable auxiliaire non modifiée x_i est:

$$\beta_{\text{MC}}(F_S) = \frac{\sum y_i/\hat{\pi}_i}{\sum x_i/\hat{\pi}_i}, \quad (10)$$

avec les poids $w_i = \eta(x'_i, y'_i) - \beta_{\text{MC}}(x'_i) / (y'_i - \beta_{\text{MC}}(x'_i))$. Il s'agit en fait du résultat de la première étape de l'algorithme par les moindres carrés repondérés par itérations, qu'on utilise souvent pour calculer les estimateurs M. L'estimateur HTR à un degré hérite de la plupart des propriétés utiles de l'estimateur HTR entièrement itéré; il est plus facile à mettre en oeuvre et plus rapide à calculer.

4. ESTIMATEURS À RISQUE ESTIMÉ MINIMUM

L'estimateur HTR est généralement biaisé. L'erreur quadratique moyenne (EQM), $E_S[(\bar{x}_U \beta(F_S, \eta) - y_U)^2]$, de l'échantillonnage constitue un critère de rendement pratique. Pour les échantillons petits à moyens, les gains obtenus avec les estimateurs HTR par rapport à l'estimateur HT ne sont pas très sensibles à la méthode d'accroissement de la robustesse retenue lorsque l'échantillon comporte des valeurs aberrantes (voir Hulliger 1991, chapitre 3). Toutefois, lorsque les données sont adéquates et que les échantillons sont moyens à grands, les pertes en matière d'EQM de certains estimateurs HTR risquent d'être énormes. La question qui se pose est celle de choisir un bon estimateur HTR. Les estimateurs à risque estimé minimum (estimateurs REM) qui adaptent la constante de mise au point d'un estimateur HTR à l'échantillon constituent une possibilité. Hulliger (1991, chapitre 2) se penche sur l'utilisation des estimateurs REM pour l'espérance d'une variable aléatoire univariée. L'idée consiste à prendre un estimateur M simple – l'estimateur M de Huber, par exemple – d'estimer l'EQM pour une série de constantes de mise au point k , et de choisir la constante de mise au point qui présente l'erreur quadratique moyenne estimée la plus petite.

3.4 Estimateurs à un degré

Il est déconseillé d'exprimer les estimateurs robustes sous forme de moyennes pondérées avec observations assorties de poids fixes puisque la notion et l'effet d'une valeur aberrante dépend du domaine particulier et de la variable à analyser. Toutefois, les estimateurs "à un degré" qui sont exprimés sous forme de moyennes pondérées réduisent la complexité des calculs des estimateurs robustes. L'estimateur HTR à un degré s'obtient par la formule

$$\bar{x}_U = \frac{\sum_{i \in S} w_i y'_i x'_i / \pi_i}{\sum_{i \in S} w_i x'_i / \pi_i}, \quad (11)$$

l'exige en principe le plan PIPT, ceci signifie que la CS de l'estimateur HT est bornée en x_i . En d'autres mots, l'estimateur HT est à l'épreuve des valeurs aberrantes x_i . Toutefois, la borne risque d'être quantitativement trop élevée pour être efficace et il pourrait s'avérer nécessaire de donner un poids moins élevé aux valeurs aberrantes x_i .

2.4 Espérance et variance

approximatives

En suivant le chemin-

Gwet et Rivest (1992), on

peut montrer que $\beta(F_S, \eta)$

est cohérent pour $\beta(F_U, \eta)$,

si on songe au fait que pour une séquence de population

hiérarchique en croissance et les échantillons PPT,

l'espérance de $\beta(F_S, \eta)$ est approximativement égale

à $\beta(F_U, \eta)$. Evidemment, $x_U \beta(F_U, \eta)$ peut être différent

de la moyenne de la population et alors, $x_U \beta(F_S, \eta)$ sera

biaisé comme estimateur de y_U . En particulier, si la distri-

bution de la population n'est pas symétrique, $x_U \beta(F_S, \eta)$

constituera en général un estimateur biaisé pour y_U même

si l'reste cohérent pour $x_U \beta(F_U, \eta)$. La question qui se

pose alors est de savoir quelle est l'ampleur du biais de

$x_U \beta(F_S, \eta)$, en particulier lorsqu'on la compare avec la

variance.

La CS (5) peut servir à dériver une approximation de

la variance. La dérivation est analogue au cas des variables

aléatoires indépendantes à distribution identique, où la

fonction d'influence est remplacée par la CS d'échantil-

lonnage. En partant de l'espérance du carré de (6), on

obtient après certaines approximations

$$\text{Var}_S \beta(F_S, \eta) \approx E_S[(\beta(F_S, \eta) - \beta(F_U, \eta))^2]$$

$$\approx \frac{\text{Var}_S(\sum_{i \in S} \eta(x_i, r_i) / \pi_i)}{(\sum_{i \in U} \eta(x_i, r_i) / \pi_i)^2}$$

$$\approx \frac{\left(\frac{1}{\pi_i} - 1 \right) \eta(x_i, r_i)^2 x_i^2 + \sum_{i \neq j \in U} \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) \eta(x_i, r_i) \eta(x_j, r_j) x_j^2}{\sum_{i \in U} \eta(x_i, r_i)^2 x_i^4 + \sum_{i \neq j \in U} \eta(x_i, r_i) \eta(x_j, r_j) x_j^2} \quad (8)$$

3.1 Stratification et domaines

3. GÉNÉRALISATION

Le signe moins dans l'équation (9) est approprié. Les produits (négatifs) croisés dans le numérateur dominent habituellement. Néanmoins, v_{rHT} risque de devenir négatif, tout comme l'estimateur de variance HT (1) (voir Cochran 1977, p.261). L'estimateur de variance v_{rHT} ne donne pas l'estimateur de variance (1) si $\eta(x, r) \equiv r$. On peut évidemment utiliser l'estimateur de Yates-Grundy-Sen pour faire l'estimation du numérateur de V_r . Un troisième estimateur de variance peut être dérivé en associant l'estimateur HTR à un estimateur par les moindres carrés pondéré dont les poids dépendent de l'estimation (voir Hülliger 1991, p. 166). Comme les estimateurs RBM (voir section 4) ont donné des résultats légèrement meilleurs avec v_{rHT} qu'avec les autres estimateurs de variance, les simulations de la section 5 ont été faites avec v_{rHT} .

Le numérateur de V_r est la variance de $\sum_{i \in S} \eta(x_i, r_i)$ ($\beta(F_U, \eta) x_i / \pi_i$, laquelle est un estimateur HT séparé de l'inconnu r_i ($\beta(F_U, \eta)$). Par conséquent, on peut utiliser l'estimateur de la variance (1) pour l'estimateur HT. Après le remplacement de $\beta(F_U, \eta)$ par l'estimateur $\beta(F_S, \eta)$, l'estimateur de la variance de l'estimateur HTR devient

$$v_{rHT} = -x_U^2 \frac{\sum_{i \in S} \frac{\pi_i}{1} \eta(x_i, r_i)^2 x_i^2 + \sum_{i \neq j \in S} \frac{\pi_{ij}}{1} \eta(x_i, r_i) \eta(x_j, r_j) x_j^2}{\sum_{i \in S} \frac{\pi_i}{1} \eta(x_i, r_i)^2 x_i^4 + \sum_{i \neq j \in S} \frac{\pi_{ij}}{1} \eta(x_i, r_i) \eta(x_j, r_j) x_j^2} \quad (9)$$

2.5 Estimation de la variance

La moyenne stratifiée d'un échantillonnage aléatoire stratifié est un estimateur HT. On peut l'assimiler à la moyenne des valeurs prévues en vertu d'un modèle d'analyse de la variance simple. La méthode correspondante d'accroissement de la robustesse est directe; elle consiste en un échantillonnage de la robustesse de chacune des moyennes des strates (Hülliger 1991). Toutefois, si la taille de l'échantillon de strate n'est que de 1 ou 2, on peut réduire le poids des valeurs aberrantes sans avoir à poser d'autres hypothèses. Par ailleurs, les biais des moyennes de strates à robustesse accrue peuvent s'additionner pour donner un biais d'ensemble important (voir Rivest 1993, section 4). Par conséquent, il faudra peut-être recourir à des méthodes d'accroissement de la robustesse différentes pour l'estimation des moyennes des strates et des moyennes générales.

où r_i est évalué à $\beta(F_U, \eta)$. Désignons cette variance approximative par V_r . Ce cas diffère surtout de celui de la variance asymptotique d'un estimateur M avec variables aléatoires indépendantes à distributions identiques du fait que les produits croisés du numérateur de V_r ne disparaissent pas. Si $\eta(x, r) \equiv r$, alors V_r donne la variance correcte de l'estimateur HT.

$(N + 1) [\lambda(\beta(F_{U+}, \eta), F_{U+}) - \lambda(\beta(F_U, \eta), F_U)] = 0$. En utilisant une approximation linéaire de $\eta(x, \cdot)$ et en négligeant les termes dans $1/N$, on peut isoler la courbe de sensibilité de $\beta(F_U, \eta)$ de l'équation suivante:

$$(N + 1) (\beta(F_{U+}, \eta) - \beta(F_U, \eta)) \approx \frac{\eta(x', r') x'}{\sum_{i \in U} \eta_2(x'_i, r'_i) x_i'^2 / N} =: \text{CS}(x, y, F_U, \eta), \quad (5)$$

où $\eta_2(x, r) = \partial \eta(x, r) / \partial r$ et où r' et r'_i sont tous deux évalués à $\beta(F_U, \eta)$. Cette courbe de sensibilité peut être étendue au cas d'une variable explicative p -dimensionnelle (voir Hampel et coll. 1986, p. 316, et Hulliger 1991, p. 183). Comme les unités ne sont habituellement pas incluses dans un échantillon PIPT de façon indépendante, la réaction de la pente HTR à une observation particulière doit être examinée en conditionnant sur un échantillon particulier. La déviation de l'estimateur $\beta(F_S, \eta)$ pour un échantillon particulier S de $\beta(F_U, \eta)$ peut être déterminée approximativement par l'intégration de la courbe de sensibilité de $\beta(F_S, \eta)$ en fonction de la fonction de distribution de l'échantillon F_S (voir Hampel et coll. 1986, p. 85) comme suit:

$$\beta(F_S, \eta) - \beta(F_U, \eta) \approx \int \text{CS}(x, y, F_U, \eta) dF_S. \quad (6)$$

L'influence de l'unité i dans l'échantillon S peut être assimilée à la contribution de l'unité i à la déviation due à l'échantillon S , c'est-à-dire:

$$\text{CS}((x_i, \pi_i, y_i) \mid S, F_U, \eta) =$$

$$\frac{\eta(x'_i, r'_i) x_i' / \pi_i}{(\sum_{j \in S} 1 / \pi_j) \sum_{j \in U} \eta_2(x'_j, r'_j) x_j'^2 / N}. \quad (7)$$

La CS peut être examinée sous un angle théorique pour traiter des propriétés de l'estimateur HTR et choisir une bonne fonction η . Elle peut en outre être estimée par remplacement du facteur de normalisation $N / (\sum_{j \in U} \eta_2(x'_j, r'_j) x_j'^2)$ par un estimateur approprié. La CS estimée peut servir à la détection des valeurs aberrantes.

L'influence de l'unité i dans l'échantillon S sur l'estimateur HT est définie par

$$x_U \text{CS}((x_i, \pi_i, y_i) \mid S, F_U, \eta) \equiv (r) =$$

$$(y_i - \beta_{MC}(F_U)(x_i)) \left/ \left(1 + \pi_i \sum_{j \in S \setminus i} 1 / \pi_j \right) \right.$$

Cette CS n'est pas bornée en y_i , de sorte que l'estimateur HT n'est pas à l'épreuve des valeurs aberrantes y_i . Le y_i exerce une influence sur l'estimateur HT par l'intermédiaire du résidu $y_i - \beta_{MC}(F_U)(x_i)$. On comprend alors clairement pourquoi une valeur y_i élevée combinée à une valeur x_i faible (ou un π_i faible) peut avoir une grande

fonctionnelle de moindres carrés. Ni l'estimateur HT ni l'estimateur HTR n'ont besoin de ce modèle ou d'une symétrie d'erreurs à variance proportionnelle à x pour pouvoir s'appliquer.

D'autres formulations de l'estimateur HT en fonctionnelles de moindres carrés peuvent s'avérer appropriées dans certaines conditions. Supposons qu'en dépit du plan PIPT, il n'y ait pas de corrélation de y_i avec π_i . Dans un tel cas, on choisirait probablement la moyenne de l'échantillon non pondérée $\bar{y}_S = \sum_{i \in S} y_i / n$ en guise d'estimateur de la moyenne de la population (voir Rao 1966). L'accroissement de la robustesse de \bar{y}_S pourrait devenir une solution ponctuelle. Si l'estimateur HT est en fait approprié à cause de la corrélation entre y_i et π_i , l'accroissement de la robustesse ne sera pas efficace.

Dans un troisième type d'accroissement de la robustesse, on présumerait que y_i est proportionnel à x_i , mais avec une variance proportionnelle au carré de x_i . Il s'agit en fait de la situation où l'estimateur HT est optimal. L'accroissement correspondant de la robustesse serait une solution β de $\sum_{i \in S} \eta(x_i, y_i / x_i - \beta) = 0$. De toute évidence, cet accroissement de la robustesse ne tient pas compte du plan d'échantillonnage PIPT. Si ce plan est réintégré dans l'équation d'estimation en résolvant $\sum_{i \in S} \eta(x_i, y_i / x_i - \beta) / \pi_i = 0$, l'estimateur HT n'est pas récupéré lorsque $\eta(x, r) \equiv r$.

On pourra prétendre qu'en fait, l'estimateur HT n'est jamais utilisé sous cette forme pure pour l'estimation des moyennes de populations. L'estimateur habituel est $(\sum_{i \in S} y_i / \pi_i) / (\sum_{i \in S} 1 / \pi_i)$, un estimateur qu'on appelle parfois estimateur Hájek. L'équation d'estimation de l'estimateur Hájek, $\sum_{i \in S} (y_i - T) / \pi_i = 0$, montre clairement que l'estimateur Hájek n'est pas à l'épreuve des valeurs aberrantes dans y . Toutefois, le résidu $y_i - T$ ne fait pas entrer en jeu la variable auxiliaire x_i . Par conséquent, l'estimateur Hájek ne souffre pas d'un possible effet combiné d'un y_i grand et d'un x_i petit, ce qui peut devenir un point en faveur du modèle de régression qui sous-tend l'estimateur HT.

2.3 Courbe de sensibilité de l'échantillonnage

Pour dériver une variance d'échantillonnage approximative de l'estimateur HTR (voir section 2.4), on utilise une population finie analogue à la fonction d'influence pour les populations infinies (Hampel 1974). Pour l'échantillonnage d'une population finie avec courbe de sensibilité le plan, il convient d'élaborer une courbe de sensibilité (CS) (voir Hampel et coll. 1986, p. 93) pour $\beta(F_U, \eta)$ à la fonction de distribution de la population F_U . En d'autres mots, la pente de la fonctionnelle HTR est linéarisée autour de F_U .

Désignons par $U+$ la population U augmentée d'une unité présentant les caractéristiques (x, y) , et désignons par $\lambda(\beta, F_U)$ la fonction $\sum_{i \in U} \eta(x'_i, r'_i) (\beta) (x'_i) / N$, de telle manière que l'équation de définition pour $\beta(F_U, \eta)$, l'estimateur M à la fonction de distribution de la population, est $\lambda(\beta, F_U) = 0$. Il apparaît alors clairement que

élaboré par Horvitz et Thompson, ou par l'estimateur de variance mis au point par Yates, Grundy et Sen (voir Cochran 1977, p. 261).

Dans la documentation spécialisée en échantillonnage d'enquête, on justifie l'emploi de l'estimateur HT par le fait qu'il a une variance d'échantillonnage égale à zéro si les probabilités d'inclusion π_i sont exactement proportionnelles à y_i . Dans ce cas, $T_{HT}(y_S) = \bar{y}_U$ pour chaque échantillon S . L'estimateur HT est robuste en ce qui concerne la distorsion, mais non en ce qui concerne la variance pour les écarts par rapport à la proportionnalité entre y_i et π_i (voir Rao 1966).

Comment pourrait-on formuler l'estimateur HT de manière à autoriser la dérivation d'une fonction d'influence analogue à d'un estimateur de variance? L'idée maîtresse consiste à exprimer l'estimateur HT en tant que fonctionnelle de moindres carrés (MC) d'une estimation de la fonction de distribution de la population, de manière que le plan soit incorporé dans l'estimateur de cette fonction pendant que la proportionnalité de y_i et x_i est incorporée dans la fonctionnelle MC.

La fonction conjointe de distribution de la population des deux variables (x_i, y_i) prend la forme $F_U(r, t) = \sum_{i \in U} \mathbf{1}\{x_i \leq r\} \mathbf{1}\{y_i \leq t\} / N$, où $\mathbf{1}\{y_i \leq t\} = 1$ si $y_i \leq t$, ou 0 dans les autres cas. Il existe divers moyens d'estimer F_U , mais l'estimateur le plus simple et le plus généralement applicable est la fonction de distribution de l'échantillon

$$F_S(r, t) = \sum_{i \in S} \frac{1}{\pi_i} \mathbf{1}\{x_i \leq r\} \mathbf{1}\{y_i \leq t\} / \sum_{i \in S} \frac{1}{\pi_i}. \quad (2)$$

L'estimateur F_S est lui-même une fonction de distribution. Pour dériver une fonctionnelle MC, on utilise le modèle suivant d'une superpopulation pour la proportionnalité entre y_i et x_i : nous présumons que y_U est un vecteur de réalisations des variables aléatoires indépendantes X_i , avec une espérance βx_i et une variance $\sigma^2 x_i$.

Définition 1. L'estimateur MC $\beta_{MC}(F_S)$ de β du modèle décrit ci-haut en ce qui a trait à la fonction de distribution de l'échantillonnage F_S de (x_i, y_i) ($i \in S$) minimise $\int (y - \beta x)^2 / x dF_S(x, y)$ ou résout de façon équivalente

$$\sum_{i \in S} \frac{1}{\pi_i} \left(y_i - \beta x_i \right) \left(\frac{y_i}{x_i} - \beta \right) = 0. \quad (3)$$

L'énoncé suivant est bien connu et facile à prouver. Si S est un échantillon tiré conformément à un plan d'échantillonnage PPT assorti de probabilités d'inclusion $\pi_i = nx_i / \sum_{i \in U} x_i$ ($i \in U$), l'estimateur HT est alors $T_{HT} = x_U \beta_{MC}(F_S)$, où $\beta_{MC}(F_S)$, et l'estimateur MC défini en (3) est donné par

$$\beta_{MC}(F_S) = \frac{\sum_{i \in S} y_i / \pi_i}{\sum_{i \in S} x_i / \pi_i}.$$

A noter que l'expression $T_{HT} = x_U \beta_{MC}(F_S) = x_U$ ($\sum_{i \in S} y_i / \pi_i$) / ($\sum_{i \in S} x_i / \pi_i$) ne dépend pas du modèle de

2.2 L'estimateur HT à robustesse accrue

Après la séparation du plan et de l'information auxiliaire et son expression sous forme de fonctionnelle MC, l'accroissement de la robustesse de l'estimateur HT devient une opération analogue à celle de l'accroissement de la robustesse des estimateurs MC dans les modèles linéaires pour des populations finies à l'aide des estimateurs M (voir Hampel et coll. 1986, chapitre 6). L'équation d'estimation (3) comporte maintenant une fonction η qui dépend des résiduelles normalisées $(y_i - \beta x_i) / x_i^{1/2}$ et de x_i . Pour simplifier l'écriture, nous désignons par (\cdot) la division par $x_i^{1/2}$ et posons que $r'(\beta) = (y - \beta x) / x^{1/2}$.

Définition 2. Soit $\beta(F_S, \eta)$ une solution de l'équation

$$\sum_{i \in S} \frac{1}{\pi_i} \eta(x'_i, r'_i(\beta)) x'_i = 0. \quad (4)$$

L'estimateur HT à robustesse accrue (estimateur HTR) est

$$T_{HTR}(F_S) : = x_U \beta(F_S, \eta)$$

où $\beta(F_S, \eta)$ est appelé la pente de l'estimateur HTR.

En général, les choix utiles de η prennent la forme de $\eta(x, r) = w(x) \psi(r \cdot u(x))$, où $w(x)$ et $u(x)$ sont deux fonctions de pondération, et ψ est une fonction de définition pour un estimateur M ponctuel (voir Hampel et coll. 1986, p. 315). Dans les calculs qui suivent, nous utilisons la "forme de Mallows" dans laquelle $u(x) \equiv 1$. La régression de type Mallows donne un poids moindre aux valeurs x aberrantes et aux valeurs résiduelles aberrantes de façon indépendante. Un exemple bien connu, dans lequel on a également $w(x) \equiv 1$, est la fonction de Huber $\eta(x, r) = \psi_{\text{Hub}}(r, k) = \max(-k, \min(k, r))$ pour une certaine constante k . L'estimateur HTR dont la fonction de définition $\eta(x, r) \equiv r$ est l'estimateur HT. Ainsi, en ajustant la constante de mise au point k dans la fonction de Huber, il est possible d'obtenir une transition graduelle des estimateurs, de l'estimateur HT à d'autres estimateurs de plus en plus robustes.

Des estimations d'échelle sont nécessaires dans $w(x)$ et $\psi(r)$ pour conférer à $\beta(F_S, \eta)$ l'équivariance d'échelle. Alors que des estimateurs d'échelle préliminaires sont disponibles pour la fonction de pondération $w(x'_i)$ (p. ex., la médiane de x'_i), l'échelle des résidus doit être estimée simultanément avec la pente β . On peut utiliser la médiane des résidus absolus. Dans le développement théorique qui suit (sections 2.3 à 4), on présume que l'échelle est connue pour simplifier le traitement.

L'estimateur HTR est un estimateur non paramétrique. Le modèle $E y = \beta x$ est utilisé essentiellement pour motiver l'expression de l'estimateur HT sous la forme d'une

d'échantillonnage aléatoire simple sans remise. Oehlert (1985) propose l'estimateur de mode à moyenne aléatoire pour l'estimation robuste de la moyenne de populations finies. Smith (1987) rappelle qu'il est tout aussi important de détecter et de traiter les observations influentes si l'inférence est fondée sur la randomisation fournie par le plan d'échantillonnage, que si les observations sont considérées comme des manifestations de variables aléatoires. Il propose une mesure d'influence pour les estimateurs linéaires fondée sur l'élimination des cas et qui fait intervenir à la fois la variable visée et son poids.

La démarche prédictive, en théorie de l'échantillonnage, s'appuie sur des modèles stochastiques de la population pour prévoir le total de la réalisation actuelle. On utilise des modèles linéaires et des estimateurs (non robustes) linéaires. Iachan (1984) se penche sur les aspects de la sensibilité et de l'accroissement de la robustesse dans les cas où les modèles sont mal définis. Chambers (1986) élabore une méthode d'accroissement de la robustesse de la prédiction qui s'appuie sur l'utilisation d'estimateurs M. Il établit une distinction entre les valeurs aberrantes représentatives et non représentatives dans un échantillon. Les valeurs aberrantes représentatives doivent être incluses avec leur poids entier dans une estimation non biaisée de la moyenne de la population, tandis que les valeurs aberrantes non représentatives doivent recevoir un poids plus faible ou être rejetées.

Little et Smith (1987) traitent les valeurs aberrantes et les valeurs manquantes dans certaines séries de données positives multivariées continues avec un algorithme EM à robustesse accrue. Gwet et Rivest (1992) se penchent sur les estimateurs par quotient robustes dans un contexte d'échantillonnage aléatoire simple sans remise. Les estimateurs M constituent une classe d'estimateurs robustes à la fois simples et souples. Un estimateur M ponctuel T est défini implicitement par l'équation d'estimation

$$\sum_{i=1}^t \psi(X_i - T) = 0$$

pour une fonction ψ déterminée à l'avance telle que $\psi_{\text{HUB}}(x, k) = \max(-k, \min(k, x))$, où k est une constante de mise au point. Un estimateur M peut se présenter comme élément d'une fonction de distribution empirique. La fonction d'influence d'un estimateur est une dérivée fonctionnelle de l'estimateur (Hampel 1974). Elle décrit la réaction de l'estimateur à une légère contamination des données. Un estimateur M à fonction ψ bornée comporte habituellement une fonction d'influence bornée qui fait en sorte que les valeurs aberrantes ne nuisent pas trop à l'estimateur. Pour les besoins de l'estimation de la moyenne des populations asymétriques finies, les estimateurs M doivent être adaptés. Dans le présent article, nous proposons des estimateurs M adaptés aux échantillons assortis de probabilités d'inclusion inégales. Le modèle linéaire simple qui constitue le fondement implicite de la stratégie de Horvitz-Thompson est rendu explicite, et l'estimateur HT est assimilé à un fonctionnel

2. ACCROISSEMENT DE LA ROBUSTESSE DES ESTIMATEURS HORVITZ-THOMPSON

2.1 L'estimateur HT en tant que fonctionnelle de moindres carrés

Une population finie $U = \{1, \dots, N\}$ de $0 < N < \infty$ éléments distincts est échantillonnée. Nous nous intéressons à la variable y qui prend les valeurs y_i pour $i \in U$. Le plan d'échantillonnage $p(S)$ dans l'espace des échantillons S d'éléments distincts assortis de probabilités d'inclusion $\pi_i = P[i \in S] = \sum_{S \ni i} p(S)$. Ces valeurs de π_i sont proportionnelles à une certaine variable auxiliaire positive connue x_i ($i \in U$). On appelle ces plans d'échantillonnage PPT (probabilité d'inclusion proportionnelle à la taille) puisque x_i est souvent une mesure de taille. Désignons par π_{ij} la probabilité d'inclusion conjointe $P[i \in S, j \in S]$ ($i, j \in U$). Le vecteur de toutes les valeurs de y prend la forme $y_U = (y_1, \dots, y_N)^T$ et x_U est défini d'une manière analogue. Le vecteur des valeurs y d'un échantillon S prend la forme $y_S = (y_1, \dots, y_n)^T$ ($i_k \in S$). L'objectif consiste à estimer la moyenne de la population de la variable y : $\bar{y}_U = \sum_{i \in U} y_i / N$. L'estimateur HT pour \bar{y}_U est $T^{HT} = \sum_{i \in S} y_i / (N\pi_i)$. La variance de T^{HT} s'obtient par l'estimateur bien connu

$$v^{HT}(T^{HT}) = \frac{1}{N^2} \left[\sum_{i \in S} (1 - \pi_i) \frac{y_i^2}{\pi_i^2} + \sum_{i \neq j \in S} (1 - \pi_i \pi_j / \pi_{ij}) \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} \right], \quad (1)$$

Estimateurs Horvitz-Thompson à l'épreuve des valeurs aberrantes

BEAT HULLIGER¹

RÉSUMÉ

L'estimateur Horvitz-Thompson (estimateur HT) résiste mal aux valeurs aberrantes. La présence de valeurs aberrantes dans une population peut accroître sa variance, même si elle demeure non biaisée. Nous exprimons l'estimateur HT sous forme d'une fonctionnelle de moindres carrés afin d'en accroître la robustesse par l'intermédiaire d'estimateurs M. Nous dérivons une variance approximative de l'estimateur HT à robustesse accrue en utilisant un type de fonction d'influence pour l'échantillonnage, et nous élaborons un estimateur de cette variance. Une méthode adaptée de sélection d'un estimateur M conduit à des estimateurs à risque estimé minimum. Ces estimateurs et les estimateurs HT à robustesse accrue sont souvent plus efficaces que les estimateurs HT classiques, lorsque nous sommes en présence de valeurs aberrantes.

MOTS CLÉS: Valeur aberrante; estimateur M; adaptation; moyenne de la population; échantillonnage; courbe de sensibilité.

1. INTRODUCTION

La moyenne d'une variable appartenant à une population finie est un indicateur important. On peut songer par exemple à la moyenne des salaires pratiqués dans un secteur donné de l'économie, ou au rendement moyen de la culture du maïs dans une région agricole. À cause de son rapport avec la somme, la moyenne n'est pas facile à remplacer par d'autres indicateurs. Cependant, la moyenne d'une population est une caractéristique sensible puisque une seule observation démesurée peut suffire à en déterminer la valeur. L'estimateur Horvitz-Thompson (estimateur HT) est un estimateur naturel de la moyenne d'une population lorsque le plan d'échantillonnage est assorti de probabilités d'inclusion inégales et que l'échantillonnage s'effectue sans remise. Il correspond à la moyenne de l'échantillon dans un échantillonnage aléatoire simple. Il n'est jamais biaisé, peu importe la distribution de la population estimée, la moyenne de la population, il s'agit d'un estimateur linéaire. Les observations démesurées, comme les faibles probabilités d'inclusion, ont une incidence particulièrement grande sur l'estimateur HT.

Supposons qu'il existe une valeur aberrante dans un échantillon. Il pourrait s'agir d'une observation correcte issue de la population cible. Le rejet d'une telle valeur rendrait l'estimateur HT biaisé. Cependant, en conservant cette valeur avec tout son poids, on rendrait l'estimateur HT extrêmement variable puisque la valeur aberrante n'apparaîtrait, typiquement, que dans quelques-uns des échantillons possibles. Il y a donc un effet de compensation entre la distorsion et la variance dans des cas comme celui-ci qui comprennent, en particulier, les distributions asymétriques dont une des extrémités est importante.

La valeur aberrante peut également être le fait d'une observation erronée, due par exemple à une erreur de

mesure ou de codage ou représentant un élément extérieur à la population cible. Dans un tel cas, la décision de conserver la valeur aberrante avec tout son poids pourrait donner un estimateur HT fortement biaisé, en plus d'accroître la variabilité. Ainsi, le rejet des valeurs aberrantes erronées a pour effet de réduire à la fois la distorsion et la variance. Comme il est souvent difficile de détecter les valeurs aberrantes et de décider si elles sont correctes ou non, il serait utile d'avoir des estimateurs qui se comportent bien, tant du point de vue de la distorsion que de celui de la variance, sans égard à la nature des valeurs aberrantes possibles et à leur détection. Les estimateurs HT à robustesse accrue grâce aux estimateurs M sont des candidats prometteurs pour cette difficile tâche.

Dans la documentation spécialisée sur l'échantillonnage d'enquête, le problème des valeurs aberrantes est souvent abordé sous la rubrique des "populations asymétriques". Kish décrit le problème qu'elles posent dans les enquêtes économiques et les enquêtes portant sur les personnes (Kish 1965; 11.4 B). Il propose la formation de strates séparées pour les valeurs aberrantes, lorsque c'est possible, ainsi que la troncature, la transformation ou la modélisation. L'idée de former des classes séparées pour les unités démesurées et de combiner les moyennes de classes est examinée, par exemple, par Glasser (1962) et par Hidiroglou et Srinath (1981).

L'idée de la troncature est précisée avec la moyenne "winsorisée" proposée par Searls (1966). Fuller (1991) suggère pour sa part un estimateur de test préliminaire ayant pour effet de réduire l'incidence des valeurs les plus grandes dans le seul cas où un test des valeurs extrêmes donne des résultats significatifs. Rívest (1993) se penche sur le comportement de divers plans de winsorisation dans un contexte d'échantillonnage aléatoire simple. Shoemaker et Rosenberger (1983) dérivent des formules exactes pour le calcul de la valeur attendu et de la variance de la médiane et de la moyenne tendancielle dans un contexte

¹ Beat Hulliger, Office fédéral de la statistique de Suisse, Schwarztortstrasse 96, CH-3003, Berne, Suisse.

- MEEDEN, G., et VARDEMAN, S. (1991). A noninformative Bayesian approach to interval estimation in finite population sampling. *Journal of the American Statistical Association*, 86, 972-980.
- MEEDEN, G. (1993). Noninformative nonparametric Bayesian estimation of quantiles. *Statistics and Probability Letters*, 16, 103-109.
- RAO, J.N.K., KOVAR, J.G., et MANTEL, H.J. (1990). On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika*, 77, 365-375.
- ROYAL, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57, 377-387.
- ROYAL, R.M., et CUMBERLAND, W.D. (1981). An empirical study of the ratio estimator and estimators of its variance. *Journal of the American Statistical Association*, 76, 66-88.
- ROYAL, R.M., et CUMBERLAND, W.D. (1985). Conditional coverage properties of finite population confidence intervals. *Journal of the American Statistical Association*, 80, 355-359.
- RUBIN, D.B. (1981). The Bayesian bootstrap. *Annals of Statistics*, 9, 130-134.

ces résultats ici puisqu'ils correspondent étroitement aux résultats de la "distribution a posteriori de Polya".) Nous voyons que leur comportement conditionnel, du moins dans ces cas, ressemble beaucoup à leur comportement inconditionnel. En bref, les estimations par intervalle de la médiane fondées sur la "distribution a posteriori de Polya" devraient avoir des propriétés "fréquentistes" raisonnables, quelle que soit la façon dont l'échantillon a été choisi, pourvu que la population respecte, de façon approximative, notre opinion que les ratios sont à peu près interchangeables.

Tableau 4

Valeur moyenne et erreur absolue moyenne pour l'estimateur ponctuel et étendue moyenne et fréquence de couverture pour un intervalle de crédibilité à .95 de la médiane fondée sur la "distribution a posteriori de Polya"

pour 500 échantillons aléatoires simples tirés de l'ensemble de la population, de la moitié "la plus petite", de la moitié "la plus grande" et du tiers "médian"

Population de l'échantillon	Taille de l'échantillon	Partie de la population d'où est tiré	Valeur	Erreur moyenne	Erreur moyenne	Fréquence de couverture
ppcites	25	population totale	.195	.0072	.041	.968
		1/2 la plus petite	.192	.0047	.033	.994
		1/2 la plus élevée	.196	.0078	.048	.988
ppcounites	30	1/2 médian	.201	.0114	.055	.922
		population totale	19.4	.220	1.46	.990
		1/2 la plus petite	18.6	.305	1.34	.942
ppcounites	30	1/2 la plus élevée	18.1	.283	1.59	.954
		1/2 médian	18.5	.252	1.35	.964
		population totale	1.24	.0072	.141	.964
ppcounites	30	1/2 la plus petite	1.24	.027	.153	.966
		1/2 la plus élevée	1.23	.020	.125	.982
		1/2 médian	1.23	.027	.139	.944
ppgammada	30	population totale	43.9	.53	2.70	.950
		1/2 la plus petite	43.8	.55	2.82	.948
		1/2 la plus élevée	44.0	.53	2.55	.940
ppgammada	30	1/2 médian	43.9	.47	2.63	.974
		population totale	43.6	2.38	11.7	.932
		1/2 la plus petite	42.2	2.69	11.6	.890
ppgammada	30	1/2 la plus élevée	45.1	2.25	11.2	.950
		1/2 médian	45.2	2.27	11.3	.936

Comme on peut le voir en étudiant les tracés de y_i/x_i par rapport à x_i et les résultats de nos simulations, le fait que les ratios y_i/x_i diminuent à mesure que x_i augmentent ne semble pas avoir beaucoup d'importance. Toutefois, il est crucial que la valeur moyenne des ratios dans la bande étroite au-dessus d'un petit intervalle de valeurs x possibles reste assez constante à mesure que nous déplaçons le petit intervalle vers la droite. Dans la figure 2, le tracé des ratios pour *ppgammada* est un exemple d'une telle représentation graphique. En fait, c'est de cette façon que la population a été construite, puisqu'elle est conforme aux hypothèses qui sous-tendent *estcd*. Dans les figures 1 et 3, nous voyons, pour *ppcounites* et pour *pplin*, que la valeur moyenne des ratios dans une bande étroite tend à diminuer à mesure

REMERCIEMENTS

Cette recherche a été partiellement supportée par la subvention SES 9201718 de la NSF.

BIBLIOGRAPHIE

- BASU, D. (1971). An essay on the logical foundations of survey sampling, part one. Dans *Foundations of Statistical Inference*. Toronto: Holt, Reinhart and Winston, 203-242.
- BERGER, J.O. (1985). *Statistical Decision and Bayesian Analysis*. New York: Springer-Verlag.
- BINDER, D.A. (1982). Non-parametric Bayesian models for samples from finite populations. *Journal of the Royal Statistical Society, Series B*, 44, 388-393.
- CHAMBERS, R.L., et DUNSTAN, R. (1986). Estimating distribution functions from survey data. *Biometrika*, 73, 597-604.
- FERGUSON, T.S. (1973). A Bayesian analysis of some non-parametric problems. *Annals of Statistics*, 1, 209-230.
- KUK, A.Y.C., et MAK, T.K. (1989). Median estimation in the presence of auxiliary information. *Journal of the Royal Statistical Society, Series B*, 51, 261-269.
- MAK, T.K., et KUK, A. (1993). A new method for estimating finite-population quantiles using auxiliary information. *The Canadian Journal of Statistics*, 21, 29-38.
- MEEDEN, G., et GHOSH, M. (1983). Choosing between experiments: applications to finite population sampling. *Annals of Statistics*, 11, 296-305.

Comme autre possibilité, nous pourrions considérer un plan de sondage plus équilibré fondé sur la stratification de la population en fonction de la variable auxiliaire. Par exemple, considérons à nouveau la population *ppgammada*, où la première strate est composée des unités comptant les cinquante valeurs les plus faibles de x_i , la deuxième strate, des unités groupant les cinquante et unième à centième plus petites valeurs de x_i , et ainsi de suite. Nous avons alors tiré 500 échantillons aléatoires stratifiés de taille cinquante ou cinq unités ont été choisis au hasard dans chaque strate. Pour ces exemples, la valeur moyenne de *estdp* était de 43.94 et l'erreur absolue moyenne, de 1.81. L'étendue moyenne de l'estimateur par intervalle correspondant était de 8.95 et la fréquence relative de couverture, de .938. Il faut remarquer que ces chiffres sont très semblables à ceux des tableaux 1 et 2, où nous avons eu recours à un échantillonnage aléatoire simple.

semblable être assez robuste par rapport à l'hypothèse d'interchangeabilité.

En pratique, il arrive souvent qu'on veuille obtenir et des estimations par intervalle et des estimations ponctuelles pour les paramètres étudiés. Dans Kuk et Mak (1989) et dans Chambers et Dunstian (1986), on suggère des méthodes qui permettent de trouver, à l'aide de la théorie asymptotique, des estimations par intervalle fondées sur les estimateurs proposés par ces auteurs. Mais dans aucun cas les auteurs n'ont trouvé d'estimateurs par intervalle. Meeden et Vardeman (1991) ont fait remarquer comment on peut trouver de façon approximative des régions de crédible à 95% approximatives fondées sur la "distribution a posteriori de Polya". Si nous posons que $q(.025)$ et $q(.975)$ représentent respectivement les .025 ième et .975 ième quantiles de la collection des médianes pour 500 populations simulées sous une "distribution a posteriori de Polya", alors $(q(.025), q(.975))$ est un intervalle de crédible à 95% approximatif. (Voir Berger 1985 pour la définition de tels intervalles.) Le tableau 3 présente l'étendue moyenne et la fréquence relative de couverture de ces intervalles. Nous constatons que, pour ces populations, les intervalles ont des propriétés "fréquentistes" raisonnables. Cela n'est peut-être pas inattendu étant donné la discussion présentée dans Meeden et Vardeman (1991). Mais, par contre, une seule des populations a été construite de façon que les ratios y_i/x_i soient interchangeables. Ces résultats laissent supposer que les estimateurs ponctuels et les estimateurs par intervalle de la médiane fondés sur la "distribution a posteriori de Polya" pour les ratios sont assez robustes relativement à l'hypothèse d'interchangeabilité et qu'ils devraient donner de bons résultats dans diverses situations. Nous parlerons plus longuement de cette question dans la section 5.

Tableau 3

Étendue moyenne et fréquence relative de couverture pour un intervalle de crédible à .95 pour la médiane fondée sur la "distribution a posteriori de Polya"
pour 500 échantillons aléatoires simples

Population	Taille de l'échantillon	Étendue moyenne	Fréquence de couverture
<i>ppcites</i>	25	.041	.968
<i>ppsals</i>	30	.141	.964
<i>ppcounties</i>	30	1.44	.994
<i>ppexpd</i>	30	2.26	.944
<i>ppgama5a</i>	30	2.70	.950
	50	2.15	.956
<i>ppgama5b</i>	30	11.67	.932
	50	8.86	.942
<i>ppgama20</i>	30	3.24	.960
	50	2.51	.964
<i>ppln</i>	30	84.8	.934
	50	65.4	.956
<i>pppskew</i>	30	15.52	.936
	50	12.00	.938

Le calcul de l'estimateur *estpp* est fondé sur l'hypothèse que les ratios de population y_i/x_i sont interchangeables. Cette hypothèse peut être décrite mathématiquement de deux manières distinctes mais liées. La première est le modèle de superpopulation déjà présenté; la seconde découle de la "distribution a posteriori de Polya", qui est fondée sur un argument bayésien pas à pas et qui donne une interprétation bayésienne non informative de l'estimateur. Cette deuxième méthode peut être utilisée quel que soit le paramètre estimé. Quand on estime la moyenne, cette méthode mène à l'estimateur de Basu, qui donne à peu près le même résultat que l'estimateur par le quotient, bien que ce dernier donne habituellement un résultat légèrement supérieur. Quand on estime la médiane, cette méthode mène à l'estimateur dont on traite dans le présent article. Ici, nous avons soutenu que la "distribution a posteriori de Polya" pour les ratios mène à de bons estimateurs ponctuels et à de bons estimateurs par intervalle de la médiane quand une variable auxiliaire est présente et qu'elle semble assez robuste par rapport à l'hypothèse de l'interchangeabilité des ratios y_i/x_i .

Royall et Cumberland (1981) ont présenté une étude empirique de l'estimateur par le quotient et d'estimateurs de sa variance. Ils ont soutenu qu'étant donné un échantillon il arrive souvent qu'une estimation de la variance, fondée sur le modèle de superpopulation, qui mène à l'estimateur par le quotient, soit plus raisonnable qu'une estimation fondée sur un plan basée sur une distribution d'échantillonnage probabiliste. Dans Royall et Cumberland (1985), ils ont démontré que, sous réserve de la moyenne de l'échantillon de la variable auxiliaire, les propriétés de couverture conditionnelle de l'intervalle de crédible habituel fondé sur un plan, pour la moyenne de la population, n'étaient absolument pas fiables.

Nous allons maintenant aborder la question du comportement conditionnel des intervalles pour la médiane fondée sur la "distribution a posteriori de Polya" qui ont été élaborés dans cet article. Dans les études de simulation présentées plus tôt, nous avons utilisé un échantillonnage aléatoire simple pour des raisons d'ordre pratique. Pour obtenir une idée du comportement conditionnel de la "distribution a posteriori de Polya", nous avons étudié cinq de nos populations. Dans chaque cas, nous avons classé la population à l'aide des valeurs de la variable auxiliaire x . Nous avons alors tiré 500 échantillons aléatoires de la première (ou plus petite) moitié de la population, puis 500 échantillons aléatoires additionnels de la deuxième (ou plus grande) moitié de la population et finalement 500 échantillons aléatoires de plus dans le tiers médian de la population. Puis, nous avons calculé l'intervalle de crédible à .95 pour la médiane fondé sur la "distribution a posteriori de Polya" qui suppose l'interchangeabilité des ratios y_i/x_i . Dans le tableau 4, nous présentons les résultats pour les estimateurs fondés sur la "distribution a posteriori de Polya" de la médiane. (Nous avons aussi calculé la valeur moyenne et l'erreur absolue moyenne de l'estimateur *estcd* pour ces exemples. Nous n'avons pas inclus

valeur moyenne de tous les estimateurs sauf *estm* est présentée. Tous les estimateurs sont approximativement sans biais sauf dans un cas, *estcd* pour la population *ppln*. Nous n'avons pas inclus les résultats pour *estsm*, puisqu'il est bien établi que cet estimateur est sans biais. Dans le tableau 2, l'erreur absolue moyenne pour les six estimateurs est présentée. Nous voyons, dans ce tableau, que *estcd* et *estdp* sont de loin les meilleurs estimateurs. Ils donnent tous les deux de meilleurs résultats que les quatre autres estimateurs dans tous les cas sauf un. Dans *ppexp*, *estsm* a donné de meilleurs résultats que ces deux estimateurs, mais il s'agit d'un cas où l'on ne pouvait s'attendre à ce que l'un ou l'autre de ces deux estimateurs donne de bons résultats. Pour les sept premières populations, le rendement était presque identique alors que, pour la population *ppln*, l'estimateur *estdp* est préféré et que, pour la population *ppstskew*, c'est le contraire.

Tableau 1

Valeur moyenne de cinq estimateurs de la médiane pour 500 échantillons aléatoires simples				
Population (médiane)	Taille de l'échantillon	Valeur moyenne de l'estimateur		
		<i>estm</i>	<i>estsm</i>	<i>estcd</i> <i>estdp</i>
<i>ppcites</i> (1.90)	25	.197	.196	.193
<i>ppscites</i> (1.24)	30	1.21	1.25	1.23
<i>ppcounties</i> (18.33)	30	18.21	18.60	18.66
<i>ppexp</i> (29.02)	30	29.03	29.05	29.00
<i>ppgmma5a</i> (43.90)	30	43.82	43.88	43.91
<i>ppgmma5b</i> (44.17)	30	43.84	43.96	44.19
<i>ppgmma20</i> (44.17)	30	44.28	44.37	44.18
<i>ppln</i> (170.25)	30	171.15	169.38	168.12
<i>ppstskew</i> (46.12)	30	43.66	40.27	45.88
	50	169.15	167.54	167.65
	50	170.61	185.01	170.61
	50	23.34	23.18	23.43
	50	23.47	23.28	23.46
	50	43.90	43.91	43.85
	50	43.89	43.99	43.90
	50	43.61	44.15	43.98
	50	23.77	23.14	23.43
	50	169.61	185.03	169.61
	50	45.11	45.50	45.43
	50	45.37	46.01	45.43

Tableau 2

Erreur absolue moyenne de six estimateurs de la médiane pour 500 échantillons aléatoires simples				
Population	Taille de l'échantillon	Erreur absolue moyenne de l'estimateur		
		<i>estm</i>	<i>estsm</i>	<i>estcd</i> <i>estdp</i>
<i>ppcites</i>	25	.0326	.0161	.0162
<i>ppscites</i>	30	.1797	.0770	.0797
<i>ppcounties</i>	30	3.12	.586	.964
<i>ppexp</i>	30	.43	.49	.48
<i>ppgmma5a</i>	30	1.36	.96	1.03
<i>ppgmma5b</i>	30	2.84	2.74	2.71
<i>ppgmma20</i>	30	2.08	2.04	2.01
<i>ppln</i>	30	25.9	25.8	24.2
<i>ppstskew</i>	30	3.86	4.26	6.69
	50	18.0	20.1	17.9
	50	17.7	16.46	17.7
	50	12.7	21.4	17.0
	50	.49	.51	.49
	50	.64	.67	.64
	50	1.85	1.80	1.89
	50	2.38	2.37	2.58
	50	.43	.44	.44
	50	.53	.54	.54
	50	.46	.48	.47
	50	.214	.215	1.34
	50	.0075	.0075	.0075
	50	.0072	.0072	.0072

Dans la majorité des exemples, où l'on utilise des estimateurs par quotient, tant les y_i que les x_i sont habituellement strictement positifs. Dans la population *ppstskew*, 13 des 1,000 unités ont une valeur y négative. Dans la construction originale de la population *ppln*, un nombre beaucoup plus important de valeurs étaient négatives. La population a été modifiée pour que toutes les valeurs soient supérieures à zéro.

On remarquera que ces populations ont été construites selon divers scénarios pour le lien entre les variables x et y . *Ppgmma20* et *ppgmma5a* respectent les hypothèses du modèle de superpopulation qui mènent à *estcd*, alors que *ppgmma5b* est compatible avec les hypothèses qui sous-tendent *estdp*. Dans *ppstskew*, la variance conditionnelle de y_i étant donné x_i , est compatible avec *estcd*, alors que pour la population *ppln* non modifiée elle était compatible avec *estdp*. Dans ces deux cas, l'hypothèse pour l'espérance conditionnelle n'est pas respectée. Dans le cas des populations *ppcounties*, *ppgmma5a* et *ppln*, nous avons représenté graphiquement y par rapport à x et y/x par rapport à x . Les résultats sont présentés dans les figures 1 à 3.

L'estimateur *estdp* est fondé sur l'hypothèse qu'étant donné l'échantillon s , notre opinion à propos des ratios observés, c.-à-d. les ratios y_i/x_i pour $i \in s$, et notre opinion sur les ratios non observés, c.-à-d. les ratios y_j/x_j pour $j \notin s$ sont à peu près interchangeables. En particulier, cela signifie que l'opinion qu'on peut avoir à propos d'un ratio y_j/x_j ne devrait pas dépendre de la taille de x_j . En fait, *ppgmma5b* a été constituée de façon que cela soit effectivement vrai. Par contre, dans le cas du modèle de superpopulation qui mène à l'estimateur *estcd*, nous nous attendrions à ce que la variabilité des ratios diminue à mesure qu'augmente la taille de la variable x tandis que la valeur moyenne des ratios dans toute bande verticale étroite demeure à peu près constante lors du déplacement de la bande vers la droite. Cela se voit clairement dans le tracé des ratios pour la population *ppgmma5a*. Pour les autres populations, sauf pour *ppgmma20*, la valeur des ratios dépend en fait de la taille de x . Cela se voit clairement dans les tracés pour *ppcounties* et *ppln*. Par conséquent, ces populations devraient constituer des cas types intéressants pour l'estimateur *estdp*. *Ppexp* a été incluse comme cas type pour qu'on puisse voir ce qui se produirait si l'on respectait peu les hypothèses sur lesquelles sont fondés *estdp* et *estcd*.

4. CERTAINS RÉSULTATS DES SIMULATIONS

Pour comparer les six estimateurs, 500 échantillons aléatoires simples de diverses tailles ont été tirés parmi les neuf populations. Pour chaque échantillon, la valeur des six estimateurs a été calculée. Pour l'estimateur *estdp*, cela signifie qu'il a fallu trouver cet estimateur de façon approximative en simulant $R = 500$ réalisations de la distribution prédictive de la médiane de la population induite par la "distribution a posteriori de Polya". Dans chaque cas, on a calculé la valeur moyenne et l'erreur absolue moyenne de l'estimateur. Dans le tableau 1, la

paramètre de forme et un comme paramètre d'échelle. Alors, étant donné x_i , la distribution conditionnelle de y_i était normale avec moyenne $3x_i$ et variance x_i^2 .

Dans *ppskew*, la variable auxiliaire était fortement désaxée vers la droite avec moyenne 42.63, médiane 39.29 et variance 204.59. Alors, étant donné x_i , la distribution conditionnelle de y_i était normale avec moyenne $x_i + 5$ et variance $9x_i$.

Dans *ppln*, la variable auxiliaire était un échantillon aléatoire provenant d'une population lognormale avec moyenne et écart-type (du logarithme) de 4.9 et de .586 respectivement. Alors, étant donné x_i , la distribution conditionnelle de y_i était normale avec moyenne $x_i + 2 \log x_i$ et variance x_i^2 .

Dans *ppexp*, la valeur de la variable auxiliaire était cinquante plus une valeur d'un échantillon aléatoire provenant de la distribution exponentielle traditionnelle. Alors, étant donné x_i , la distribution conditionnelle de y_i était normale avec moyenne $80 - x_i$ et variance $(.6 \log x_i)^2$.

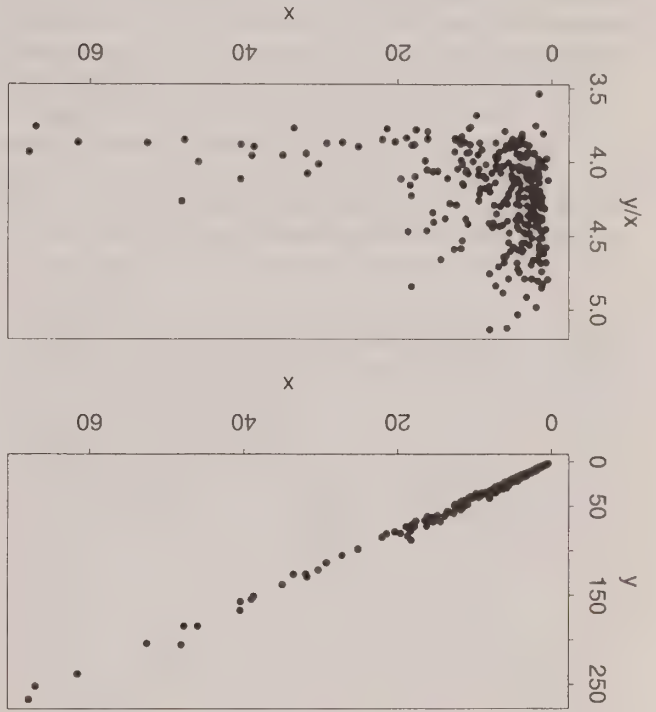


Figure 1. Pour *ppcounties*, tracé de y par rapport à x et de y/x par rapport à x où x représente le nombre de familles (en milliers) vivant dans un comté et y , la population totale (en milliers) du comté pour 304 comtés.

Toutes les populations renferment 500 unités sauf *ppskew*, qui en a 1,000. Les coefficients de corrélation entre les deux variables pour ces six dernières populations sont de .76, .87, .41, .61, .58 et -.28, respectivement.

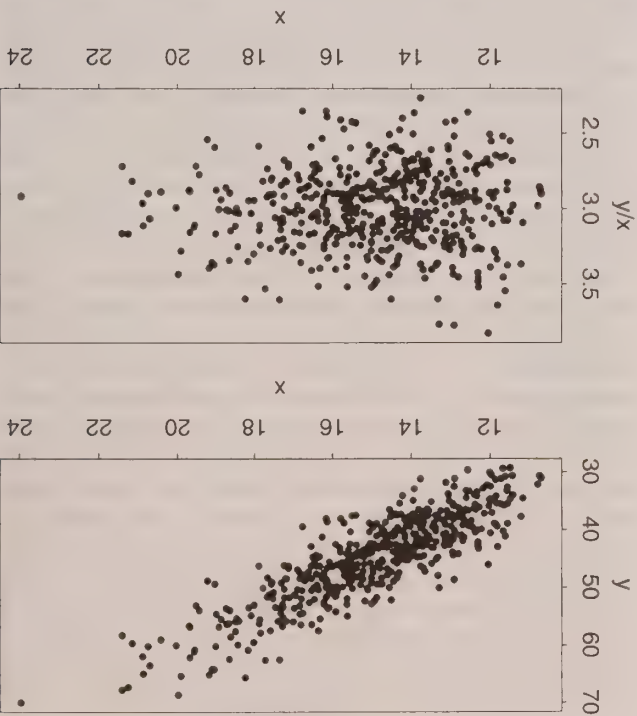


Figure 2. Pour *ppgamm5a*, tracé de y par rapport à x et de y/x par rapport à x .

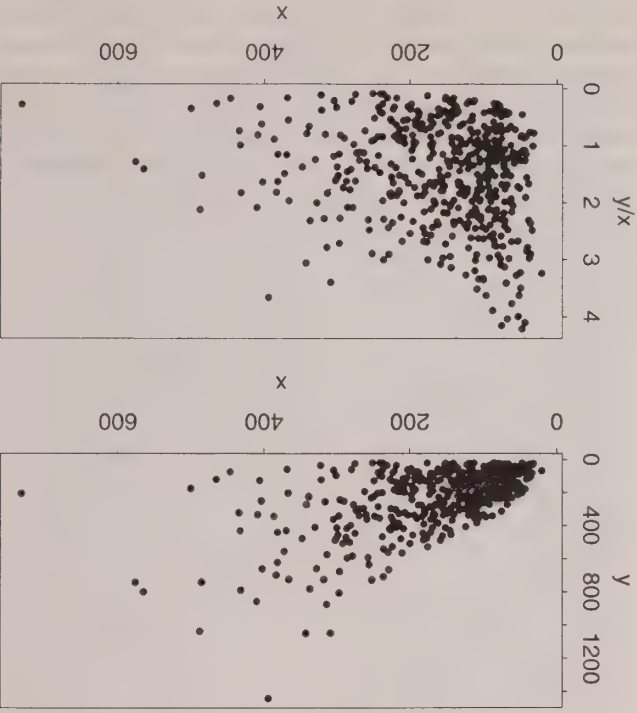


Figure 3. Pour *ppln*, tracé de y par rapport à x et de y/x par rapport à x .

pourrait agir comme si les n ratios connus ($y_i - b x_i$)/ $\sqrt{x_i}$ pour $i = 1$ à s sont des observations réelles tirées de cette distribution inconnue. Selon cette hypothèse, pour un t fixe et une unité j fixe qui ne fait pas partie de l'échantillon s , une estimation de $\Delta(t - y_j)$ est donnée tout simplement par le nombre des n ratios connus qui incorporent b inférieurs ou égaux à $(t - b x_j)/\sqrt{x_j}$ divisé par n . Finalement, si nous calculons la somme, pour toutes les unités j non observées, des estimations de $\Delta(t - y_j)$, nous avons donc une estimation de la seconde somme dans l'expression ci-dessus pour $F(t)$ qui donne alors une estimation de $F(t)$. Une fois que nous pouvons estimer $F(t)$ pour tout t par, disons, $\hat{F}(t)$, l'estimation de la médiane de la population est $\inf\{t: \hat{F}(t) \geq 0.5\}$.

3. LES POPULATIONS

Nous allons comparer ces estimateurs à l'aide de plusieurs populations différentes. Nous commençons avec trois populations réelles. La première est un groupe de 125 villes américaines. La variable x est leur population en 1960, en millions d'habitants, alors que leur variable y est la population correspondante en 1970, à nouveau, en millions d'habitants. La deuxième est un groupe de 304 comtés américains. La variable x est le nombre de familles dans les comtés en 1960, alors que la variable y est la population totale du comté en 1960. Les deux variables sont présentées en milliers d'habitants. La troisième population est un groupe de 331 grosses sociétés. La variable x est leurs ventes totales en 1974 et la variable y , leurs ventes totales en 1975. Les ventes sont présentées en milliards de dollars. Nous représentons ces trois populations par *ppcites*, *ppcounties* et *ppsals*. Pour les trois populations, les coefficients de corrélation sont .947, .998 et .997, respectivement. Ces populations ont été étudiées dans Royall et Cumberland (1981). Notre *ppcounties* ressemble à leur population Counties60 sauf que, pour nous, la variable x représente le nombre de familles plutôt que le nombre de ménages.

Nous avons aussi étudié six populations artificielles. Dans chaque cas, nous avons d'abord choisi la variable auxiliaire x , puis nous avons produit la variable y à partir de la variable auxiliaire. Dans certains cas, nous avons suivi le modèle de superpopulation décrit au début de la section précédente pour un certain choix des u_i . Dans d'autres cas, nous n'avons pas respecté l'hypothèse selon laquelle la moyenne de y_i est $b x_i$, étant donné la valeur x_i . Dans tous les cas, les erreurs, les e_i étaient des variables aléatoires indépendantes suivant une distribution normale identique de moyenne nulle et de variance un. Dans la première population, *ppgamma20*, les x_i étaient un échantillon aléatoire tiré d'une distribution gamma avec vingt comme paramètre de forme et un comme paramètre d'échelle. Alors, étant donné x_i , la distribution conditionnelle de y_i était normale avec moyenne $1.2 x_i$ et variance x_i , c.-à-d. $u_i = \sqrt{x_i}$. Dans la deuxième population, *ppgamma5a*, la valeur des x_i était de dix plus une valeur d'un échantillon aléatoire provenant d'une distribution gamma avec cinq comme

population. Selon le plan de sondage de Polya, pour les ratios décrits ci-dessus, nous pouvons simuler une réalisation possible de toute la population. Pour cette copie simulée, nous pouvons alors trouver sa médiane. Si nous répétons ce processus R fois, nous aurons alors simulé la distribution prédictive de la médiane de la population selon la "distribution a posteriori de Polya". Quand R est grand, la moyenne de ces R médianes simulées de la population donne, approximativement, l'estimateur *esdp*. Nous allons maintenant comparer l'estimateur *esdp* à plusieurs autres estimateurs, dont un est tout simplement la médiane de l'échantillon des y_i . Pour cet estimateur, représenté par *esism*, que nous employons comme étalon, nous ne tenons pas compte des renseignements contenus dans la variable auxiliaire. Un autre estimateur est l'analogie naturel de l'estimateur par le quotient de la moyenne de la population. Cet estimateur est étudié dans Kuk et Mak (1989), et nous le représentons par *estrm*. Il ne s'agit que du ratio de la médiane des valeurs y sur la médiane des valeurs x de l'échantillon, multiplié par la médiane de toutes les valeurs x dans la population. Les auteurs ont proposé deux autres estimateurs pour la médiane. Nous n'étudierons que le premier d'entre eux et le représenterons par *estkm*. Cet estimateur a une justification intuitive plausible qui est présentée dans leur article. Rao, Kovar et Mantel (1990) ont étudié un estimateur fondé sur un plan de la médiane. Nous représenterons cet estimateur par *estkm*. Comme le calcul de cet estimateur peut prendre beaucoup de temps, nous en obtiendrons une approximation à l'aide d'une méthode présentée dans Mak et Kuk (1993). Enfin, nous allons considérer l'estimateur proposé dans Chambers et Dunstan (1986) et le représenter par *esicd*. En fait, Chambers et Dunstan proposent toute une famille d'estimateurs, et nous n'en étudierons qu'un cas spécial qui convient lorsque $u_i = \sqrt{x_i}$ dans le modèle de superpopulation décrit au début de la présente section. Nous allons maintenant présenter brièvement le raisonnement qui les amène à leur estimateur de la médiane. Représentons par F la fonction de distribution cumulative associée aux valeurs y de la population. C'est-à-dire que F attribue un poids $1/N$ à chaque y_i de la population totale. La première étape consiste à obtenir un estimateur de $F(t)$ pour un nombre réel arbitraire t . Si s représente notre échantillon de taille n , alors, compte tenu de l'échantillon, nous pouvons poser

$$F(t) = N^{-1} \left\{ \sum_{i \in s} \Delta(t - y_i) + \sum_{j \in s} \Delta(t - y_j) \right\}$$

où $\Delta(z)$ est la fonction en escalier qui a la valeur un lorsque $z \geq 0$ et zéro ailleurs. Puisque la première somme dans l'expression ci-dessus est connue une fois que nous avons observé l'échantillon, pour obtenir une estimation de $F(t)$, il suffit de trouver une estimation de la seconde somme. Or, d'après notre modèle supposé de superpopulation, les ratios de population $(y_i - b x_i)/\sqrt{x_i}$ sont des variables aléatoires indépendantes qui suivent la même distribution. Puisque, une fois l'échantillon s observé, $b = \sum_{i \in s} y_i / \sum_{i \in s} x_i$ constitue une estimation naturelle de b , on

Pour plus de précision, supposons que notre échantillon renferme les n premières unités de la population. Nous construisons une urne qui contient n boules où l'on donne à la boule i la valeur du i -ième ratio observé, disons r_i . Nous commençons par choisir une boule au hasard dans l'urne, et la valeur observée est attribuée à l'unité non observée $n + 1$. Cette boule ainsi qu'une autre boule est choisie dans l'urne, et sa valeur est attribuée à l'unité non observée $n + 2$. Cette boule ainsi qu'une autre boule ayant la même valeur sont remises dans l'urne. On poursuit ce processus jusqu'à ce qu'un ratio ait été attribué à toutes les unités non observées. Une fois que nous avons attribué une valeur à toutes ces unités, nous avons observé une réalisation provenant de notre distribution "à posteriori" pour les ratios non observés, étant donné l'échantillon de ratios observés. Si, dans ce processus, on a attribué à l'unité non observée j le ratio avec valeur r_j , nous disons alors que sa valeur y_j est x_j . Par conséquent, à l'aide d'un échantillonnage de Polya simple, étant donné l'échantillon, nous avons créé une distribution prédictive pour les unités non observées. Nous appelons cette distribution prédictive la "distribution à posteriori de Polya". Il est facile de vérifier que cette distribution prédictive donne l'estimateur de Basu lorsque nous estimons la moyenne de la population quand nous avons une fonction quadratique de perte.

Étant donné l'échantillon, la "distribution à posteriori de Polya" donne une distribution prédictive pour les unités non observées de la population et, par conséquent, elle donne aussi une distribution prédictive pour la médiane. Du point de vue de la théorie de la décision, la fonction de perte habituelle est l'erreur absolue quand nous estimons une médiane. Pour cette fonction de perte, l'estimation bayésienne n'est que la médiane de la distribution à posteriori ou prédictive pour la médiane de la population. Si l'on utilisait une fonction quadratique de perte pour estimer la médiane, alors l'estimation bayésienne ne serait que la moyenne de la distribution prédictive pour la médiane de la population. L'admissibilité de ces estimateurs quand nous utilisons une fonction de perte appropriée découle d'un argument bayésien pas à pas de la même façon que la preuve de l'admissibilité, pour l'estimateur de Basu, de la moyenne de la population. Dans Meeden et Vardeman (1991) et Meeden (1993), on a remarqué le fait un peu surprenant décrit ci-après. Pour beaucoup de distributions courantes, la moyenne de la distribution prédictive de la médiane de la population a donné de meilleurs résultats que la médiane de la distribution prédictive de la médiane de la population selon les deux fonctions de perte. Des résultats semblables valent pour ce problème. Par conséquent, notre estimateur sera la moyenne de la distribution prédictive de la médiane de la population, bien que nous suivrions l'usage courant et utiliserions l'erreur absolue comme fonction de perte. Nous désignerons cet estimateur par *esdp*. Cet estimateur ne peut être trouvé de façon explicite. Toutefois, nous le trouverons de façon approximative en simulant des observations à partir de la distribution à posteriori ou prédictive de la médiane de la population.

Dans Meeden et Vardeman (1991), on a élaboré une approche bayésienne non informative à l'échantillonnage de populations finies, fondée sur la "distribution à posteriori de Polya". Pour le cas simple où aucune variable auxiliaire n'est présente, compte tenu des valeurs observées dans l'échantillon, cette approche introduit une distribution de Polya comme une pseudo-distribution à posteriori pour les unités non observées de la population. Cette pseudo-distribution à posteriori peut être utilisée pour obtenir des estimations ponctuelles et des estimations par intervalle de diverses quantités relatives à la population. Elle est liée à la méthode bootstrap bayésienne de Rubin (1981) et à la distribution *a priori* du processus Dirichlet de Ferguson (1973). Quand on estime la médiane, cette méthode donne des résultats semblables à ceux obtenus par Binder (1982). Un argument de Bayes pas à pas qui donne l'admissibilité des estimateurs résultants constitue une justification théorique de cette méthode. Voyez, par exemple, Meeden et Ghosh (1983). C'est dans cet article que l'on a démontré l'admissibilité de l'estimateur de Basu. Dans ce cas, on a montré que l'estimateur de Basu provenait d'une "distribution à posteriori" qui traite les ratios connus et inconnus, $r_i = y_i/x_i$ comme interchangeables. Il faut remarquer que cela ressemble beaucoup, en esprit, à la justification à l'aide du modèle de superpopulation présenté plus haut, où les ratios $r_i = y_i/x_i$ étaient indépendants et suivaient une distribution identique. Nous verrons que l'argument de Bayes pas à pas, qui sous-tend l'estimateur de Basu pour la moyenne, peut s'appliquer de façon simple aux estimateurs ponctuels et aux estimateurs par intervalle pour la médiane. Malheureusement, cela n'est pas le cas pour certains des autres estimateurs. Le ratio de la médiane des valeurs y dans l'échantillon sur la médiane des valeurs x dans l'échantillon multiplié par la médiane des valeurs x dans la population constitue un estimateur naturel, mais peut-être élémentaire, qui imite, dans un certain sens, l'estimateur par le quotient de la moyenne. Il n'y a pas de théorie connue fondée sur un modèle qui sous-tende cet estimateur, comme c'est le cas pour l'estimateur par le quotient de la moyenne.

Dans l'approche bayésienne de l'échantillonnage de populations finies, il faut préciser une distribution *a priori*. Puis, étant donné un échantillon, des inférences sont fondées sur la distribution à posteriori, qui est la distribution prédictive des unités non observées de la population compte tenu des unités dans l'échantillon. Dans l'approche bayésienne pas à pas, étant donné l'échantillon, on a toujours une distribution "à posteriori", mais elle ne découle pas d'une distribution *a priori* unique. Toutefois, on peut utiliser cette distribution "à posteriori" selon la méthode bayésienne habituelle pour trouver les estimateurs ponctuels et les estimateurs par intervalle des paramètres étudiés. Nous allons maintenant montrer comment le modèle bayésien pas à pas qui donne l'estimateur de Basu pour la moyenne peut aussi être utilisé quand on estime la médiane. Dans ce modèle, étant donné l'échantillon, la distribution prédictive des ratios non observés traite les ratios observés et les ratios non observés comme "interchangeables".

Estimation de la médiane à l'aide d'informations supplémentaires

GLEN MEEDEN¹

RÉSUMÉ

On étudie le problème de l'estimation de la médiane d'une population finie quand une variable auxiliaire est présente. On propose des estimateurs ponctuels et des estimateurs par intervalle fondés sur une approche bayésienne non informative. L'estimateur ponctuel est comparé à d'autres estimateurs possibles et l'on constate qu'il donne de bons résultats dans diverses situations.

MOTS CLÉS: Enquête par sondage; estimation; médiane; variable auxiliaire; quantile; approche bayésienne non informative.

1. INTRODUCTION

On a beaucoup étudié le problème de l'estimation de la moyenne d'une population en présence d'une variable auxiliaire dans les ouvrages sur l'échantillonnage de populations finies. L'estimateur par le quotient a souvent été utilisé dans ce cas. Pour le problème de l'estimation de la médiane d'une population, la situation est très différente. Ce n'est que depuis peu que l'on étudie ce problème. Chambers et Dunstan (1986) ont proposé une méthode pour estimer la fonction de distribution d'une population ainsi que les quantiles associés. Ils ont supposé que la valeur de la variable auxiliaire était connue pour toutes les unités dans la population et leur estimateur provenait d'une méthode fondée sur un modèle. Rao et coll. (1990) ont proposé des estimateurs par le quotient et des estimateurs de différence pour la médiane à l'aide d'une méthode fondée sur un plan. Kuk et Mak (1989) ont proposé deux autres estimateurs de la médiane de la population. Pour utiliser les valeurs mateurs de Kuk et Mak, il suffit de connaître les valeurs de la variable auxiliaire pour les unités de l'échantillon ainsi que sa médiane pour toute la population. L'efficacité de ces estimateurs dépend directement de la probabilité de "concordance" plutôt que de la validité d'une hypothèse de linéarité entre la variable étudiée et la variable auxiliaire. Récemment, Meeden et Vardeman (1991) ont traité d'une approche bayésienne non informative de l'échantillonnage de populations finies. Cette nouvelle approche utilise la "distribution a posteriori de Polya" comme distribution prédictive pour les unités non observées de la population une fois l'échantillon observé. L'approche donne souvent des estimations ponctuelles et par intervalle qui ressemblent beaucoup à celle que l'on obtient dans le cadre de la théorie "fréquentiste" classique. De plus, l'approche est facile à appliquer dans le cas de problèmes difficiles à traiter avec la théorie classique. Dans le présent article, nous montrons comment cette méthode peut être utilisée dans le cas du problème de l'estimation de la médiane d'une population quand une variable auxiliaire est présente, et nous comparons cette méthode à certaines des autres méthodes proposées.

¹ Glen Meeden, School of Statistics, University of Minnesota, Minneapolis, MN 55455, U.S.A.

Considérons une population finie contenant N unités. Pour l'unité i , représentons par y_i la caractéristique étudiée et par x_i la variable auxiliaire. Nous supposons que tant y_i que x_i sont des nombres réels et que cela est connu pour toutes les unités dans la population. Représentons par s un échantillon typique de taille n choisi par échantillonnage aléatoire simple sans remise. Pour des raisons d'ordre pratique, nous supposons un échantillonnage aléatoire simple. Avant de considérer le problème de l'estimation de la médiane d'une population, nous examinons certains faits bien connus à propos du problème de l'estimation de la moyenne. Considérons le modèle de superpopulation où l'on suppose que pour chaque i , $y_i = bx_i + u_i e_i$. Ici, b est un paramètre inconnu alors que les u_i sont des constantes connues et que les e_i sont des variables aléatoires indépendantes suivant la même distribution avec espérance mathématique nulle. Comme la moyenne de population peut être représentée par $N^{-1}(\sum_{i \in s} y_i + \sum_{j \in s} y_j)$, nous pourrions penser que $N^{-1}(\sum_{i \in s} y_i + b \sum_{j \in s} x_j)$ serait une estimation raisonnable de la moyenne chaque fois que b est une estimation raisonnable de b . Un choix particulier de b est l'estimateur par les moindres carrés pondéré où les poids sont déterminés par les u_i . Par exemple, si, pour tout i , $u_i = 1/x_i$, l'estimateur résultant n'est que l'estimateur par le quotient habituel. Alors que si, pour tout i , $u_i = x_i$, alors $b = n^{-1} \sum_{i \in s} (y_i/x_i)$ et l'estimateur résultant est celui qui a été étudié par Basu (1971). (Voir aussi Royall (1970).) À l'aide de ce modèle de superpopulation, il est facile de créer des populations où l'estimateur par le quotient a une erreur quadratique moyenne plus petite que celle de l'estimateur de Basu et inversement. Une étude de simulation assez limitée portant sur diverses populations a permis de constater que celui de l'estimateur de Basu est très semblable à celui de l'estimateur par le quotient, bien que dans la majorité des cas l'estimateur par le quotient donne un meilleur résultat que l'estimateur de Basu. Cela n'est pas inattendu, compte tenu de l'utilisation répandue de l'estimateur par le quotient.

2. ESTIMATION DE LA MÉDIANE

REMERCIEMENTS

HÄJEK, J. (1959). Optimum strategy and other problems in probability sampling. *Casopis Pro Pěstování Matematiky*, 84, 387-423.

HORVITZ, D.G., et THOMPSON, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.

MADOW, W.G., et MADOW, L.H. (1944). On the theory of systematic sampling, I. *Annals of Mathematical Statistics*, 15, 1-24.

MURTHY, M.N. (1967). *Sampling Theory and Methods*. Calcutta: Statistical Publishing Society.

MURTHY, M.N., et RAO, T.J. (1988). Systematic sampling with illustrative examples. Dans *Handbook of Statistics*. (Éds. P.R. Krishniah et C.R. Rao), (Vol. 6). Amsterdam: North-Holland, 147-185.

RAO, J.N.K. (1975). On the foundations of survey sampling. Dans *A Survey of Statistical Design and Linear Models*. (Éd. J.N. Srivastava). Amsterdam: North-Holland, 489-505.

RAO, J.N.K., et BELLHOUSE, D.R. (1978). Optimal estimation of a finite population mean under generalized random permutation models. *Journal of Statistical Planning and Inference*, 2, 125-141.

SEDRANSKY, J. (1969). Some elementary properties of systematic sampling. *Skandinavisk Aktuarietidskrift*, 1-2, 39-47.

TAYLOR, H.M., et KARLIN, S. (1984). *An Introduction to Stochastic Modeling*. Orlando: Academic Press.

WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

Cette recherche a reçu l'appui du Soil Conservation Service du U.S. Department of Agriculture, en vertu du SCS Cooperative Agreement No. 68-3A75-2-64. Jeff Goebel (Soil Conservation Service) et Wayne Fuller (Iowa State University) ont élaboré le plan MC pour l'Etat de l'Alaska. L'auteur tient à remercier les arbitres anonymes pour les commentaires constructifs formulés à l'égard d'une version antérieure de cet article.

BIBLIOGRAPHIE

BELLHOUSE, D.R. (1988). Systematic sampling. Dans *Handbook of Statistics*. (Éds. P.R. Krishniah et C.R. Rao), Vol. 6. Amsterdam: North-Holland, 125-145.

BELLHOUSE, D.R., et RAO, J.N.K. (1975). Systematic sampling in the presence of a trend. *Biometrika*, 62, 694-697.

CHANDRA, K.S., SAMPATH, S., et BALASUBRAMANI, G.K. (1992). Markov sampling for finite populations. *Biometrika*, 79, 210-213.

COCHRAN, W.G. (1946). Relative accuracy of systematic and stratified random samples for a certain class of populations. *Annals of Mathematical Statistics*, 17, 164-177.

COCHRAN, W.G. (1977). *Sampling Techniques*, 3ième Ed. New York: Wiley.

Tableau 2

Le ratio de la variance probable liée au plan en vertu d'un plan MC et de la variance probable liée au plan en vertu du plan SY pour une superpopulation constituée d'une tendance (droite de pente β_1 ou onde sinusoïdale avec période optimale, où p^* est une fonction des paramètres de la superpopulation. Le ratio du meilleur plan réalisable dans chaque ligne (si ce n'est pas SY) est en italique.

Modèle	Plans à chaînes de Markov									
	ϕ	$G_{1/2}$	ST	BA	AK	G_p^*	(p^*)			
Droite + AR	-0,5	0,2332	0,2085	0,2001	0,1666	0,2056	0,2001	(1,0000)	(1,0000)	(1,0000)
	0,0	0,2220	0,1983	0,1903	0,1821	0,1957	0,1903	(1,0000)	(1,0000)	(1,0000)
Droite + AR	0,1	0,2187	0,1950	0,1871	0,1825	0,1921	0,1871	(1,0000)	(1,0000)	(1,0000)
	0,5	0,1922	0,1702	0,1645	0,1754	0,1659	0,1645	(1,0000)	(1,0000)	(1,0000)
Droite + AR	0,9	0,0980	0,0778	0,0742	0,0768	0,0762	0,0742	(1,0000)	(1,0000)	(1,0000)
	-0,5	0,4504	0,4328	0,4262	0,3647	0,4304	0,4262	(1,0000)	(1,0000)	(1,0000)
Droite + AR	0,0	0,4344	0,4172	0,4114	0,4054	0,4153	0,4114	(1,0000)	(1,0000)	(1,0000)
	0,1	0,4291	0,4121	0,4065	0,4085	0,4094	0,4065	(1,0000)	(1,0000)	(1,0000)
Droite + AR	0,5	0,3853	0,3727	0,3724	0,4116	0,3667	0,3719	(0,8320)	(0,8320)	(0,8320)
	0,9	0,1876	0,1835	0,1914	0,2170	0,1848	0,1821	(0,5223)	(0,5223)	(0,5223)
Droite + AR	-0,5	0,9233	0,9190	0,9163	0,7941	0,9175	0,9163	(1,0000)	(1,0000)	(1,0000)
	0,0	0,9201	0,9177	0,9169	0,9160	0,9174	0,9169	(1,0000)	(1,0000)	(1,0000)
Droite + AR	0,1	0,9191	0,9175	0,9175	0,9349	0,9156	0,9174	(0,8156)	(0,8156)	(0,8156)
	0,5	0,9160	0,9289	0,9439	1,0606	0,9185	0,9135	(0,1997)	(0,1997)	(0,1997)
Droite + AR	0,9	0,8621	0,9787	1,0725	1,2710	1,0017	0,7888	(0,0981)	(0,0981)	(0,0981)
	-0,5	0,9978	0,9956	0,9935	0,8617	0,9942	0,9935	(1,0000)	(1,0000)	(1,0000)
AR pur	0,0	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	(1,0000)	(1,0000)	(1,0000)
	0,1	1,0009	1,0019	1,0028	1,0228	1,0001	1,0000	(0,0000)	(0,0000)	(0,0000)
AR pur	0,5	1,0179	1,0357	1,0536	1,1852	1,0245	1,0000	(0,0000)	(0,0000)	(0,0000)
	0,9	1,2517	1,4380	1,5814	1,8798	1,4734	1,0000	(0,0000)	(0,0000)	(0,0000)
Sinus + AR	-0,5	0,9929	0,9906	0,9884	0,8578	0,9892	0,9884	(1,0000)	(1,0000)	(1,0000)
	0,0	0,9947	0,9946	0,9945	0,9950	0,9946	0,9945	(1,0000)	(1,0000)	(1,0000)
Sinus + AR	0,1	1,0110	1,0285	1,0462	1,1775	1,0173	0,9977	(0,0364)	(0,0364)	(0,0364)
	0,5	1,0110	1,0285	1,0462	1,1775	1,0173	0,9977	(0,0364)	(0,0364)	(0,0364)
Sinus + AR	-0,5	0,6747	0,6634	0,6586	0,6008	0,6604	0,6586	(1,0000)	(1,0000)	(1,0000)
	0,0	0,6603	0,6499	0,6464	0,6770	0,6477	0,6464	(1,0000)	(1,0000)	(1,0000)
Sinus + AR	0,1	0,6554	0,6455	0,6425	0,6863	0,6421	0,6425	(1,0000)	(1,0000)	(1,0000)
	0,5	0,6149	0,6133	0,6196	0,7320	0,6041	0,6121	(0,5079)	(0,5079)	(0,5079)
Sinus + AR	-0,5	0,0668	0,0384	0,0287	0,1101	0,0323	0,0287	(1,0000)	(1,0000)	(1,0000)
	0,0	0,0656	0,0372	0,0275	0,1115	0,0311	0,0275	(1,0000)	(1,0000)	(1,0000)
Sinus + AR	0,1	0,0652	0,0368	0,0271	0,1115	0,0307	0,0271	(1,0000)	(1,0000)	(1,0000)
	0,5	0,0622	0,0339	0,0247	0,1054	0,0277	0,0245	(1,0000)	(1,0000)	(1,0000)
Sinus + AR	0,9	0,0529	0,0247	0,0154	0,1016	0,0187	0,0154	(1,0000)	(1,0000)	(1,0000)
	-0,5	0,0668	0,0384	0,0287	0,1101	0,0323	0,0287	(1,0000)	(1,0000)	(1,0000)

Quand $\beta_1 = 0$ et $\phi \neq 0$, $\xi_{(\beta, \phi)}$ est un cas spécial du modèle ξ_5 . Pour $\phi > 0$, on peut conclure aussi bien du résultat 9 que des données du tableau que le plan SY est le plus efficace, car il produit l'échantillon le plus "étale", possible; toutefois, dans le cas d'une autocorrélation faible, les autres plans MC sont des solutions valables. Le plan BA est très médiocre pour ce modèle, car le plan assure que les éléments d'une paire sur deux R_i, R_{i+1} ne seront pas éloignés de plus de a unités. (Pour la même raison, BA convient bien à une population à autocorrélation négative.) Les plans AK, $G_{1/2}$ et $G_{1/2}$ l'emportent sur le plan ST, car chacun de ces plans favorise des transitions entre états de longneur approximativement égale à a . On obtient des résultats semblables pour le modèle de superpopulation

5. DISCUSSION

La classe des plans à chaînes de Markov a été définie, et nous avons vu qu'elle incluait, comme cas spéciaux, l'échantillonnage systématique, l'échantillonnage aléatoire simple stratifié et l'échantillonnage systématique compensé. Certains nouveaux plans ont été présentés (G_p , AK), et nous avons montré que leur efficacité se comparait à celle des plans courants d'échantillonnage à une unité par strate en vertu de divers modèles de superpopulation. Les nouveaux plans fonctionnent bien, notamment, dans des exemples numériques reposant sur des superpopulations constituées d'une tendance et d'erreurs autocorrélées. C'est le genre de population auquel on s'intéresse dans de nombreux problèmes d'échantillonnage géographique, comme le National Resources Inventory de 1992 en Alaska. Un plan MC bidimensionnel mis en oeuvre pour cette enquête montre que les plans MC unidimensionnels pourraient être utilement étendus au contexte de l'échantillonnage géographique, bien que des recherches additionnelles s'imposent dans ce domaine.

D'autres travaux sur l'estimation de la variance liée aux plans MC sont également nécessaires. Comme il s'agit de plans d'échantillonnage à une unité par strate, l'estimation sans biais liée au plan de la variance de l'estimateur de Horvitz-Thompson n'est pas possible. Le problème de l'estimation de la variance pour les plans à une unité par strate, en particulier le plan SY, a fait l'objet d'abondants travaux. Par exemple, Wolter (1985) examine en détail huit estimateurs biaisés de la variance pour SY et évalue leurs biais en vertu de modèles de superpopulation. Des travaux semblables, portant sur l'estimateur de la variance de strates groupées (p. ex. Cochran 1977, p. 139) en vertu de plans MC généraux, sont en cours.

$$\xi_{(\alpha, \phi)} : y_{ij} = \alpha \sin \frac{2\pi j}{d} + e_{ij}$$

où le terme central est monotone croissant à mesure que $\rho \in [0, 1]$ décroît. Si n est pair, le terme de gauche de (3) est égal à $\min_{MC} E_{\xi_3} [V_{MC}(t_\pi)]$.

3.4 Modèle de population périodique

Dans le cas d'une population affichant une périodicité déterministe (avec période égale à p), on peut appliquer le modèle simple de l'onde sinusoidale

$$\xi_4: y_{ij} = \alpha \sin \left\{ \frac{p}{2\pi} [(i-1)a + j] \right\} + e_{ij},$$

où les e_{ij} sont des variables aléatoires non corrélées de moyenne zéro et de variance σ^2 .

Résultat 6 En vertu du modèle de population périodique ξ_4 ,

$$E_{\xi_4} [V_{MC}(t_\pi)] = a^2 \alpha^2 V_{MC} \left[\sum_{i=1}^n \sin \frac{p}{2\pi} [(i-1)a + R_i] \right] + na(a-1)\sigma^2$$

pour n importe quel plan MC.

Désignons le modèle de l'onde sinusoidale ξ_4 avec $p = a$ par ξ_{4a} . En vertu de ξ_{4a} ,

$$\left\{ \sin \left\{ \frac{p}{2\pi} [(i-1)a + j] \right\} = \sin \frac{a}{2\pi} j \right\}$$

de sorte que le modèle est additif et qu'aucun plan MC n'a une variance probable liée au plan supérieure à celle du plan SY. On voit donc que le plan SY ne convient pas à une population affichant une périodicité dont la période est égale à l'intervalle d'échantillonnage (Madaw et Madaw 1994). Ce résultat peut être généralisé comme suit.

Résultat 7 Si $\mu_{ij} = \beta_j$ dans ξ , alors ξ est un modèle qui s'applique à une population affichant une périodicité déterministe, avec période égale à l'intervalle d'échantillonnage, a . Le modèle ξ est additif, de sorte qu'en vertu de ξ , aucun plan MC n'a une variance probable liée au plan supérieure à celle du plan SY.

3.5 Modèle autocorrélé

De nombreux auteurs, depuis Cochran (1946), ont comparé les plans ST et SY, ainsi que l'échantillonnage aléatoire simple, en vertu d'un modèle de superpopulation autocorrélé. Voir Bellhouse (1988, §4) pour un bilan. Considérons le modèle d'autocorrélation suivant, dû à Cochran (1946):

$$\xi_5: y_{ij} = \mu + e_{ij}$$

où $\sigma_{i',i''} = \gamma [(i' - i)a + j' - j]$ pour $i' \geq i$. **Résultat 8** En vertu du modèle autocorrélé ξ_5 ,

$$E_{\xi_5} [V_{MC}(t_\pi)] = na(a-1)\gamma(0) - 2n \sum_{h=1}^a \gamma(h)(a-h) + 2a \sum_{a=1}^{n-1} \sum_{a=1}^h \gamma(ha+j' - j)(n-h) \left(P_{ff}^{(h)} - \frac{1}{a} \right)$$

pour n importe quel plan MC.

Résultat 9 Si, pour $h \geq 0$, $\gamma(h)$ est non négatif, non croissant et convexe, c.-à-d.

$$\gamma(h) \geq 0, \gamma(h) \geq \gamma(h+1) \quad \text{et}$$

$$\gamma(h+2) - 2\gamma(h+1) + \gamma(h) \geq 0,$$

alors $E_{\xi_5} [V_{SY}(t_\pi)] = \min_{MC} E_{\xi_5} [V_{MC}(t_\pi)]$.

Ce résultat est un corollaire d'un théorème dû à Hájek (1959), donné comme théorème 4.1 dans Bellhouse (1988); Bellhouse a clarifié les conditions dans lesquelles le théorème est valable. Le théorème de Hájek généralisait un résultat antérieur de Cochran (1946), qui comparait les plans SY et ST, ainsi que l'échantillonnage aléatoire simple.

4. EFFICACITÉ: QUELQUES EXEMPLES NUMÉRIQUES

Une importante classe de modèles s'appliquant à des processus temporels et spatiaux est décrite ainsi: tendance polynomiale d'ordre peu élevé, plus séquence d'erreurs autocorrélés. On peut donner, comme exemple simple,

$$\xi_{(\beta,\phi)}: y_{ij} = \beta_0 + \beta_1 [(i-1)a + j] + e_{ij},$$

où la structure d'autocorrélation est celle du modèle autorégressif (AR) de premier ordre,

$$\sigma_{i',i''} = \gamma [(i' - i)a + j' - j] = \sigma^2 \phi^{(i' - i)a + j' - j}$$

pour $i' \geq i$ et $|\phi| < 1$. La variance moyenne liée au plan en vertu de ce modèle est obtenue des résultats 4 et 8. Pour différents choix de β_1 et ϕ , le ratio des variances probables liées au plan,

$$(4) \quad E_{\xi} [V_{MC}(t_\pi)] / E_{\xi} [V_{SY}(t_\pi)],$$

est donné au tableau 2 pour divers plans MC. Le tableau comprend également les valeurs pour le plan G_p optimal, obtenues en minimisant (4) par rapport à ρ . L'utilisation de ce plan n'est possible que si les paramètres de la superpopulation sont connus; il est donc présenté comme repère plutôt que comme choix possible.

Quand $\beta_1 \neq 0$ et $\phi = 0$, le modèle est ξ_3 et les valeurs du tableau concordent avec le résultat 5: le plan SY est le pire plan MC et le plan BA est le meilleur, tandis que les plans $G_{1/2}$, $G_{2/3}$ et ST se situent entre les deux. Bien que le plan BA soit extrêmement performant pour ce modèle, tout plan MC autre que le plan SY serait un bon choix.

$$V^{SY} \left[\sum_{i=1}^n \mu_{iR_i} \right] = V^{SY} \left[\sum_{i=1}^n \alpha_i + n\beta_{R_1} \right] = n^2 V(\beta_{R_1}),$$

tandis qu'en vertu d'un plan MC général, ce terme est

$$V^{MC} \left[\sum_{i=1}^n \mu_{iR_i} \right] = \sum_{i=1}^n \sum_{i'=1}^n C^{MC}(\beta_{R_i}, \beta_{R_{i'}}).$$

Puisque $C^{MC}(\beta_{R_i}, \beta_{R_{i'}}) \leq V(\beta_{R_1})$, la proposition est démontrée. □

Certains modèles particuliers sont examinés dans les cinq sous-sections suivantes.

3.1 Modèle de permutation aléatoire

Dans le cas d'une population en ordre aléatoire, on peut

appliquer un modèle de permutation, dans lequel une réalisation des mesures y_1, \dots, y_N est donnée par l'une des $N!$ permutations également probables de N valeurs fixes. Ce modèle peut s'écrire ainsi

$$\xi_1 : y_{ij} = y^U + e_{ij},$$

où $y^U = \sum y^k/N$. Voir Rao (1975) pour plus de détails. Le résultat suivant est alors une conséquence du théorème 2.1 de Rao et Bellhouse (1978).

Résultat 2 En vertu du modèle de permutation aléatoire,

$$E_{\xi_1} [V^{MC}(t_\pi^*)] =$$

$$(N^2/n)(1 - n/N) \sum (y_k - \bar{y})^2 / (N - 1)$$

pour n'importe quel plan MC.

Ainsi, la variance moyenne pour l'ensemble des permutations est exactement $V^{SI}(t_\pi^*)$, où SI dénote l'échantillonnage aléatoire simple (non stratifié) sans remise. Dans le cas de l'échantillonnage SY, on doit ce résultat originellement à Madow et Madow (1944). Voir aussi Sedransk (1969).

3.2 Modèle d'effets de stratification

Dans le cas d'une population comportant des effets de stratification, on peut appliquer le modèle suivant

$$\xi_2 : y_{ij} = \alpha_i + e_{ij},$$

où les α_i sont des constantes fixes et les e_{ij} sont des variables aléatoires non corrélées de moyenne zéro et de variance σ^2 . Notons que si $\alpha_i \equiv \mu$, ξ_2 est une solution de remplacement à ξ_1 comme modèle d'une population en ordre aléatoire.

Résultat 3 En vertu d'un modèle d'effets de stratification,

3.3 Modèle de tendance linéaire

pour n'importe quel plan MC.

$$E_{\xi_2} [V^{MC}(t_\pi^*)] = na(a - 1)\sigma^2$$

Dans le cas d'une population qui présente une tendance linéaire, on peut appliquer le modèle suivant

$$\xi_3 : y_{ij} = \beta_0 + \beta_1[(i - 1)a + j] + e_{ij},$$

où β_0 et β_1 sont des constantes fixes et les e_{ij} sont des variables aléatoires non corrélées $(0, \sigma^2)$.

Résultat 4 En vertu du modèle de tendance linéaire ξ_3 ,

$$E_{\xi_3} [V^{MC}(t_\pi^*)] = \beta_1^2 a^2 V^{MC} \left[\sum_{i=1}^n R_i \right] + na(a - 1)\sigma^2 \quad (2)$$

pour n'importe quel plan MC. Puisque ξ_3 est additif, aucun plan MC n'a une variance probable supérieure à celle du plan SY en vertu d'un modèle de tendance linéaire. Le seul terme lié au plan dans (2) est $V^{MC}[\sum_{i=1}^n R_i]$. Selon le plan SY, $\sum_{i=1}^n R_i = nR_1$, de sorte que

$$V^{SY} \left[\sum_{i=1}^n R_i \right] = n^2 V(R_1),$$

tandis que selon le plan ST,

$$V^{ST} \left[\sum_{i=1}^n R_i \right] = nV(R_1).$$

Selon le plan BA, si n est pair,

$$V^{BA} \left[\sum_{i=1}^n R_i \right] = V^{BA} \left[\frac{2}{n} R_1 + \frac{2}{n} (a + 1 - R_1) \right] = 0.$$

Il en découle que si la population est parfaitement linéaire ($\sigma^2 = 0$),

$$E_{\xi_3} [V^{BA}(t_\pi^*)] = 0,$$

de sorte que $t_\pi^* = t$ pour tous les échantillons, comme l'a signalé Murthy (1967, p. 165).

Résultat 5 En vertu du modèle de tendance linéaire ξ_3 ,

$$E_{\xi_3} [V^{BA}(t_\pi^*)] \leq E_{\xi_3} [V^{ST}(t_\pi^*)]$$

$$\leq E_{\xi_3} [V^{G_d}(t_\pi^*)]$$

$$\leq E_{\xi_3} [V^{SY}(t_\pi^*)] = \max_{MC} E_{\xi_3} [V^{MC}(t_\pi^*)], \quad (3)$$

3. ESTIMATION DE HORVITZ-THOMPSON EN VERTU D'UN PLAN MC

Ecrivons le total de la population comme suit

$$t = \sum_U y_k = \sum_n \sum_a y^{(t-1)a+j} = \sum_a \sum_n y_{ij}.$$

Pour tout k les probabilités d'inclusion de premier ordre d'un plan MC sont données par

$$\pi_k = \Pr\{k \in s\} = \Pr\{R_i = j\} = 1/a$$

et pour $k \leq l$, les probabilités d'inclusion de second ordre sont données par

$$\pi_{kl} = \begin{cases} 1/a, & \text{pour } i = i', j = j', \\ 0, & \text{pour } i = i', j \neq j', \\ P_{(i'-i)/a}^{ff}, & \text{pour } i > i'. \end{cases}$$

L'estimateur de Horvitz-Thompson non biaisé selon le plan (Horvitz et Thompson 1952) pour le total de la population est alors

$$\hat{t}_\pi = \sum y_k / \pi_k = a \sum_n \sum_a y_{ij} I_{\{R_i=j\}},$$

où

$$I_{\{R_i=j\}} = \begin{cases} 1, & \text{si } R_i = j, \\ 0, & \text{si } R_i \neq j. \end{cases}$$

Les covariances liées au plan des indicateurs $I_{\{R_i=j\}}$ sont données par

$$\begin{aligned} C_{MC}(I_{\{R_i=j\}}, I_{\{R_{i'}=j'\}}) &= E_{MC}[I_{\{R_i=j\}} I_{\{R_{i'}=j'\}}] - \\ &= E_{MC}[I_{\{R_i=j\}}] E_{MC}[I_{\{R_{i'}=j'\}}] \\ &= \pi_{(i-1)a+j, (i'-1)a+j'} - \\ &= \pi_{(i-1)a+j} \pi_{(i'-1)a+j'}, \end{aligned} \quad (1)$$

de sorte que la variance liée au plan de \hat{t}_π est

$$+ a^2 \sum_n \sum_a \sum_{j' \neq j} \sum_{i' \neq i} \left[\frac{a}{1} - \frac{a}{2} \right] y_{ij} y_{i'j'}.$$

pour tous les plans MC.

$$E_{\hat{t}}[V_{MC}(\hat{t}_\pi)] \geq E_{\hat{t}}[V_{SY}(\hat{t}_\pi)]$$

où $E_{\hat{t}}[e_{ij}] = 0$, $V_{\hat{t}}(e_{ij}) = \sigma_{ij}^2$ et $C_{\hat{t}}(e_{ij}, e_{i'j'}) = 0$. Alors

$$\hat{\xi} : y_{ij} = \mu_{ij} + e_{ij} = \alpha_i + \beta_j + e_{ij},$$

Proposition 2 Considérons un modèle additif non corrélé, moyenne liée au plan pire que celle du plan SY.

La proposition suivante donne une condition suffisante en vertu de laquelle aucun plan MC n'a une variance pendant de j , on a $V_{MC}[\sum_{i=1}^n \mu_{iR_i}] = 0$.

pour n'importe quel plan MC. Notons que si μ_{ij} est indé-

$$\begin{aligned} &+ 2a \sum_n \sum_a \sum_{i' > i} \sum_{j' \neq j} \sigma_{ij, i'j'} \left[P_{(i'-i)}^{ff} - \frac{a}{1} \right] \\ &(a-1) \sum_a \sum_{j=1}^n \sigma_{ij}^2 - \sum_a \sum_{j=1}^n \sum_{j' \neq j} \sigma_{ij, ij'} \end{aligned}$$

$$E_{\hat{t}}[V_{MC}(\hat{t}_\pi)] = a^2 V_{MC} \left[\sum_n \mu_{iR_i} \right] +$$

Proposition 1 En vertu du modèle de superpopulation $\hat{\xi}$, la variance moyenne liée au plan de l'estimateur de Horvitz-Thompson est

où les μ_{ij} sont fixes et les e_{ij} sont des variables aléatoires pour lesquelles $E_{\hat{t}}[e_{ij}] = 0$, $V_{\hat{t}}(e_{ij}) = \sigma_{ij}^2$ et $C_{\hat{t}}(e_{ij}, e_{i'j'}) = \sigma_{ij, i'j'}$. On peut alors utiliser, comme base de comparaison des plans, la moyenne pour le modèle de la variance liée au plan.

$$\hat{\xi} : y_{ij} = \mu_{ij} + e_{ij},$$

Puisque la variance liée au plan dépend de toutes les valeurs de la variable étudiée dans la population finie, il n'est pas facile d'utiliser (1) pour comparer les plans. À l'exemple de Cochran (1946), supposons que les valeurs de la variable étudiée sont produites à partir du modèle de superpopulation

$$\begin{aligned} &+ 2a^2 \sum_n \sum_a \sum_{i' > i} \sum_{j' \neq j} \left[P_{(i'-i)}^{ff} - \frac{a}{1} \right] y_{ij} y_{i'j'} \\ &+ 2a^2 \sum_n \sum_a \sum_{i' > i} \sum_{j' \neq j} \left[P_{(i'-i)}^{ff} - \frac{a}{2} \right] y_{ij} y_{i'j'}. \end{aligned}$$

Preuve Il découle de la proposition 1 que le seul terme qui nous intéresse est $V_{MC}[\sum_{i=1}^n \mu_{iR_i}]$ lequel en vertu d'un plan SY est

ce qui, conjugué à la propriété de Markov, entraîne que R_1, \dots, R_n sont indépendants du point de vue probabiliste. Dans ce cas, le plan MC correspond à l'échantillonnage aléatoire simple stratifié à une unité par strate (ST). **Echantillonnage systématique.** Si la matrice de probabilités de transition est I, c, a -à-d. la matrice identité $a \times a$, on a

$$\Pr\{R_i = j' \mid R_i = j\} = \begin{cases} 1, & j = j', \\ 0, & j \neq j', \end{cases}$$

de sorte qu'il existe un lien déterministe entre R_1, \dots, R_n . Par conséquent,

$$s = \{R_1, a + R_1, \dots, (n - 1)a + R_1\},$$

et le plan MC correspond à l'échantillonnage systématique (SY).

Plans mixtes. Intuitivement, on peut concevoir ST et SY comme étant, d'une certaine façon, deux "extrêmes" qui s'opposent. Si $p \in [0, 1]$,

$$G_p = p H + (1 - p) I$$

est doublement stochastique. Si $p = 0$, le plan est SY et si $p = 1$, le plan est ST. Tout autre choix de p donnera une séquence formée de "bouts" d'échantillons systématiques. Par conséquent, la classe G_p comprend les plans ST et SY, ainsi qu'un ensemble continu de plans MC mixtes. D'autres combinaisons convexes de matrices doublement stochastiques pourraient être envisagées. La classe des matrices doublement stochastiques est également fermée à l'égard de la multiplication et de la transposition des matrices, ainsi que de la permutation des lignes et des colonnes, de sorte qu'il existe de nombreuses façons de créer des plans MC.

Echantillonnage systématique compensé. Murthy (1967, §5.9d) décrit une méthode de prélèvement d'une unité par strate qu'il appelle *échantillonnage systématique compensé* ("balanced systematic sampling") (BA). Cette méthode donne les échantillons

$$s = \{R_1, a + (a + 1 - R_1), \dots, (n - 2)a + R_1,$$

$$(n - 1)a + (a + 1 - R_1)\}$$

si n est pair et

$$s = \{R_1, a + (a + 1 - R_1), \dots, (n - 2)a +$$

$$(a + 1 - R_1), (n - 1)a + R_1\}$$

si n est impair. Une caractéristique intéressante de ce plan est que si n est pair et que la population est parfaitement linéaire ($y_j = \beta_0 + \beta_1[(i - 1)a + j]$) la moyenne de l'échantillon est égale à la moyenne de la population pour n importe quel échantillon. Avec la matrice de probabilités de transition

Plan du NRI en Alaska. Comme il est indiqué à la section 1, le plan d'échantillonnage du NRI de 1992 pour la région nord-ouest de l'Alaska utilisait deux chaînes de Markov indépendantes pour la sélection contrôlée des cellules dans le sens de la latitude et dans le sens de la longitude. La matrice des probabilités de transition pour les cellules dans le sens de la longitude, R_{long} , est donnée au tableau 1. Ce plan, dénoté ci-après AK, est un plan MC, car R_{long} est doublement stochastique. La plupart des probabilités de transition sont voisines de 0.10, de sorte que la plupart des "déplacements" sont à peu près également probables. Notons toutefois que la masse a été réduite sur la diagonale inverse et à proximité, au profit des angles supérieur gauche et inférieur droit, afin que R_{long} défavorise les vastes déplacements est-ouest, par exemple de la cellule un à la cellule dix, ainsi que les courts déplacements, par exemple de la cellule dix à la cellule un. Par contre, R_{long} favorise les déplacements qui sont environ de longueur dix, par exemple de la cellule deux à la cellule un, deux ou trois. L'échantillon réalisé des cellules dans le sens de la longitude est donc bien réparti d'est en ouest, comme le serait un échantillon systématique, mais sa composante aléatoire offre une protection contre l'erreur systématique. De même, la chaîne de Markov régissant le prélèvement des cellules dans le sens de la latitude a été conçue de manière à produire une bonne dispersion géographique du nord au sud.

Tableau 1

Matrice de probabilités de transition pour l'échantillon à chaînes de Markov de cellules dans le sens de la longitude, pour le National Resources Inventory de 1992, en Alaska. Les entrées sont les probabilités conditionnelles de sélection de la cellule j' de la strate $i' + 1$ quand la cellule j de la strate i a été sélectionnée.

Cellule j de la strate i		Cellule j' de la strate $i' + 1$									
		1	2	3	4	5	6	7	8	9	10
1	0.05	0.15	0.15	0.15	0.15	0.15	0.10	0.10	0.10	0	0
2	0.15	0.15	0.15	0.15	0.10	0.10	0.10	0.10	0.10	0.05	0
3	0.15	0.15	0.10	0.10	0.10	0.10	0.10	0.05	0.05	0.10	0.10
4	0.15	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.05	0.10	0.10
5	0.15	0.10	0.10	0.10	0.10	0.05	0.05	0.10	0.10	0.10	0.15
6	0.15	0.10	0.10	0.10	0.10	0.05	0.05	0.10	0.10	0.10	0.15
7	0.10	0.10	0.05	0.05	0.10	0.10	0.10	0.10	0.10	0.10	0.15
8	0.10	0.10	0.05	0.05	0.10	0.10	0.10	0.10	0.10	0.15	0.15
9	0	0.05	0.10	0.10	0.10	0.10	0.10	0.15	0.15	0.15	0.15
10	0	0	0	0	0	0	0	0	0	0.15	0.05

Comment ce plan spécial se compare-t-il à des plans plus courants d'échantillonnage à une unité par strate? On constate, comme nous le décrirons à la section 2, qu'une application simple des chaînes de Markov permet de décrire une vaste classe de plans d'échantillonnage avec probabilités égales pour la sélection d'une unité par strate dans une population finie. Cette classe comprend des techniques courantes comme l'échantillonnage aléatoire simple stratifié, l'échantillonnage systématique et l'échantillonnage systématique compensé, ainsi que les plans appliqués à l'Alaska dont nous venons de parler. Il est facile, par ailleurs, de créer de nouveaux plans appartenant à cette classe. Ce traitement unitif des plans à une unité par strate permet de faire des comparaisons d'efficacité.

Les probabilités d'inclusion de premier et de second ordre pour tous ces plans sont obtenues à la section 3, ce qui donne l'estimateur de Horvitz-Thompson et sa variance.

Comme dans bon nombre des publications sur le sujet (Madow et Madow 1944; Cochran 1946; Sedransk 1969; Bellhouse et Rao 1975; Wolter 1985; Bellhouse 1988; etc.), la moyenne de la variance liée au plan de l'estimateur de Horvitz-Thompson est évaluée pour divers modèles de superpopulation. Des expressions compactes des moyennes, pour un modèle, des variances liées au plan sont obtenues. Des exemples numériques présentés à la section 4 montrent que les plans proposés dans cet article peuvent être plus efficaces que les plans d'échantillonnage habituels à une unité par strate pour des superpopulations qui sont la somme d'une tendance et d'erreurs autocorrélées. Un bilan est présenté à la section 5.

Bien que l'exemple qui nous inspire soit bidimensionnel, ce sont des plans unidimensionnels qui seront examinés tout au long de l'article. La plupart des démonstrations et deductions sont directes et sont omises par souci de concision.

2. PLANS À CHAÎNES DE MARKOV

Considérons le problèmes du prélèvement d'un échantillon dans une population finie de $N = na$ unités échantillonnées, désignées par

$$U = \{1, \dots, N\}$$

$$= \{1, \dots, a, a + 1, \dots, 2a, \dots,$$

$$(n - 1)a + 1, \dots, na\}.$$

La valeur d'une variable étudiée $y_k = y_{(i-1)a+j} = y_{ij}$ est associée à chaque étiquette k ; la notation y_k ou y_{ij} sera utilisée aussi bien pour les variables aléatoires que pour les réalisations des variables aléatoires.

Ici, n est la taille de l'échantillon et a est l'intervalle d'échantillonnage. Les n sous-ensembles

$$\{(i - 1)a + 1, \dots, (i - 1)a + a\} \quad (i = 1, \dots, n)$$

seront appelées les *strates*. Le but est de prélever une unité par strate. Souvent, on définit un plan d'échantillonnage

$$s = \{R_1, a + R_2, \dots, (n - 1)a + R_n\},$$

où R_1, \dots, R_n est la chaîne de Markov définie par P et $R_1 \sim \text{uniforme}(1, \dots, a)$. Formellement, donc, un *plan à chaîne de Markov* (MC) est une fonction $p(\cdot, P)$ telle que

$$p(s; P) = \Pr\{s = \{r_1, a + r_2, \dots, (n - 1)a + r_n\}\}$$

$$= \Pr\{R_1 = r_1, R_2 = r_2, \dots, R_n = r_n\}$$

$$= \begin{cases} P_{r_{n-1}, r_n} P_{r_{n-2}, r_{n-1}} \dots P_{r_1, r_2} / a, & \text{pour } r_1, \dots, r_n \in \{1, \dots, a\}, \\ 0, & \text{autrement.} \end{cases}$$

Les plans MC définis dans le présent article sont reliés aux plans présentés dans Chandrasekhar et Balasubramani (1992), dans lesquels un vecteur $1 \times N$ de probabilités initiales de sélection et une matrice $N \times N$ de probabilités de transition de périodicité n déterminent un plan d'échantillonnage sans remise. Chandrasekhar et coll. se concentrent sur la production de plans comportant des probabilités d'inclusion de second ordre strictement positives. Ils ne se penchent pas explicitement sur les plans à une unité par strate examinés ici, mais on peut directement intégrer ces derniers à leur structure, en construisant le vecteur de probabilités initiales et la matrice de probabilités de transition qui conviennent.

Le résultat suivant est utile à la détermination des caractéristiques probabilistes des plans MC.

Résultat 1

Considérons une chaîne de Markov pour laquelle la matrice de probabilités de transition P est doublement stochastique (c.-à-d. que la somme des entrées de toutes les lignes et de toutes les colonnes est égale à un) et R_1 a une distribution uniforme discrète, avec masse $1/a$ sur chacun des états $1, \dots, a$. Alors, R_i a une distribution uniforme discrète sur les états $1, \dots, a$ pour tout i . En particulier, R_i a comme moyenne $(a + 1)/2$ et comme variance $V(R_i) = (a^2 - 1)/12$.

Certains cas spéciaux de plans MC sont intéressants. **Echantillonnage aléatoire simple stratifié.** Si la matrice de probabilités de transition est

$$H = [1/a]_{j,j'=1}^a,$$

alors

$$\Pr\{R_{i'} = j' \mid R_i = j\} = 1/a = \Pr\{R_{i'} = j'\}$$

$$(j, j' = 1, \dots, a; i < i'),$$

Plans à chaînes de Markov pour l'échantillonnage à une unité par strate

F. JAY BREIDT¹

RÉSUMÉ

Dans le domaine de l'échantillonnage de populations finies, les résultats classiques indiquent que l'échantillonnage systématique est le plan le plus efficace pour l'échantillonnage à une unité par strate avec probabilités égales dans le cas de certains types de superpopulations autocorrélées, mais que l'échantillonnage aléatoire simple stratifié est peut-être nettement supérieur à l'échantillonnage systématique si la superpopulation correspond à une tendance avec erreurs non corrélées. Que faire si la superpopulation est la somme d'une tendance et d'erreurs autocorrélées? Intuitivement, on peut penser qu'un "compromis" entre les deux plans serait la meilleure solution. Dans le présent article, nous construisons de tels plans mixtes, et nous montrons qu'il s'agit d'exemples de plans à chaînes de Markov, une vaste classe de méthodes de sélection d'une unité par strate dans une population finie. Ces plans comprennent, comme cas spéciaux, l'échantillonnage systématique, l'échantillonnage systématique compensé et l'échantillonnage aléatoire simple stratifié avec une unité d'échantillonnage par strate. Les probabilités d'inclusion de premier et de second ordre sont obtenues pour les plans à chaînes de Markov, ce qui donne l'estimateur de Horvitz-Thompson et sa variance. L'efficacité de l'estimateur de Horvitz-Thompson est évaluée à l'aide de modèles de superpopulation. Des exemples numériques montrent que les nouveaux plans examinés ici peuvent être plus efficaces que les plans habituels dans le cas de superpopulations qui sont la somme d'une tendance et d'erreurs autocorrélées. Nous présentons un exemple d'application de plans à chaînes de Markov, pour les fins du National Resources Inventory de 1992 en Alaska.

MOTS CLÉS: Échantillonnage systématique compensé; National Resources Inventory; échantillonnage systématique.

L'échantillonnage stratifié, qui consiste à diviser une population finie en strates non chevauchantes et à tirer des échantillons de chaque strate, est une technique courante et efficace pour réduire l'erreur d'échantillonnage. En pratique, les plans à échantillonnage stratifié dans lesquels on prélève une seule unité d'échantillonnage par strate sont très répandus. Mentionnons par exemple l'échantillonnage aléatoire simple stratifié, ainsi que l'échantillonnage systématique et ses variantes (p. ex. Murthy et Rao 1988). Les échantillons systématiques sont vulnérables aux erreurs systématiques. Dans les échantillonnages géographiques à grande échelle, par exemple, les routes, les lignes électriques, les systèmes d'irrigation, etc. peuvent être la cause d'erreurs systématiques. Pour illustrer ce danger, on cite souvent le cas des routes délimitant des sections ("section roads"), qui existent dans certaines régions des États-Unis couvertes par l'inventaire public des terres. Ce système à quadrillage comprend des parcelles carrées appelées sections, qui ont chacune un mille de côté et qui, souvent, sont bornées par des routes dans les régions agricoles du Midwest. Le prélèvement d'un échantillon systématique avec intervalle d'un mille, si l'origine choisie au hasard tombait au mauvais endroit, pourrait laisser croire que l'Iowa est entièrement couvert de routes de gravier! L'échantillonnage systématique offre l'avantage de l'efficacité quand la population échantillonnée est positivement autocorrélée, ce qui est souvent le cas dans les problèmes d'échantillonnage temporel ou géographique,

1. INTRODUCTION

car il force les observations à être le plus possible éloignées les unes des autres, et donc le moins corrélées possible. L'autocorrélation et l'erreur systématique sont deux sujets de préoccupation dans le National Resources Inventory (NRI), un échantillonnage aréolaire des terres non fédérales aux États-Unis, réalisé tous les cinq ans par le Soil Conservation Service du United States Department of Agriculture. Les éléments de données du NRI, recueillis à la fois par télédétection et par des observations au sol, comprennent les caractéristiques du sol, l'utilisation des terres, les pratiques agricoles, les mesures d'érosion, etc. Le plan d'échantillonnage du NRI de 1992 pour la région nord-ouest de l'État de l'Alaska est une version contrôlée d'un échantillonnage à une unité par strate. La région a été divisée en bandes de latitude de vingt minutes. Chaque bande a été divisée en strates de 500,000 acres, et chacune de ces dernières a été divisée à son tour en une grille 10×10 de cellules indexées selon la latitude et la longitude. Une cellule par strate a été incluse dans l'échantillon. Dans chaque bande de vingt minutes, les strates étaient parcourues d'est en ouest pour la sélection des cellules. Les nombres aléatoires qui déterminaient les cellules sélectionnées dans le sens de la longitude, ainsi que les nombres aléatoires qui les déterminaient dans le sens de la latitude, évoluaient comme deux chaînes de Markov indépendantes. (Les résultats de base de la théorie des chaînes de Markov utilisés dans le présent article sont exposés dans des ouvrages d'introduction aux processus stochastiques, comme celui de Taylor et Karlin 1984). Les détails du plan sont présentés à la section 2.

¹ F. Jay Breidt, Iowa State University, Department of Statistics, Ames, IA 50011-1210, U.S.A.

Si les spécifications de la variance du modèle sont inexactes, de sorte qu'il y ait une certaine hétéroscédasticité, ou si la distribution est très étendue à ses extrêmes, alors la distribution des résidus le sera aussi, et les tests ne seront pas absolument stricts. Il faut prendre un certain soin pour remarquer la présence de valeurs aberrantes signalant l'existence de cette hétéroscédasticité, par exemple les flots isolés en raison d'erreurs de géocodage à grande échelle. Il y a deux façons de ne pas respecter l'hypothèse d'observations approximativement indépendantes pour l'analyse de variance. Premièrement, les UPE ne sont pas choisis par échantillonnage aléatoire simple mais plutôt à la suite d'une stratification géographique un peu plus détaillée que le plan de stratification a posteriori ne le laisse supposer. Dans la mesure où cette stratification géograp- phique réduit la variance d'échantillonnage des estimations de l'effet de l'Etat, les inférences obtenues en fonction du modèle de l'indépendance seront plutôt prudentes. Deuxièmement, il y aura des corrélations entre les facteurs de correction pour différentes parties d'ilot provenant d'un même ilot (dans les modèles comportant plusieurs GSP). Ces corrélations tendront à rendre un peu moins strictes les inférences pour lesquelles on suppose l'indé- pendance. À tout prendre, nous jugeons utiles les tests effectués dans le cadre des recherches dont les résultats sont présentés ici.

BIBLIOGRAPHIE

- ALHO, J.M., MULRY, M.H., WURDEMAN, K., et KIM, J. (1991). Estimating heterogeneity in the probabilities of enumeration for dual-system estimation. *Journal of the American Statistical Association*, 88, 1130-1136.
- BUREAU OF THE CENSUS (1990). Sample Selection Procedures for Performing Evaluation Study P12. STSD 1990 Coverage Studies and Evaluation Memorandum Series No. N-1, note de service de D. Bateman à L. Iskow et M. Lynch, 3 octobre, 1990.
- BUREAU OF THE CENSUS (1991). Request for Block Split Level Data for Performing PES Evaluation Project P12. STSD 1990 Coverage Studies and Evaluation Memorandum Series No. N-2, note de service de J. Thompson à A. Jackson, 30 janvier, 1991.
- COMMITTEE ON NATIONAL STATISTICS, PANEL TO EVALUATE ALTERNATIVE CENSUS METHODS (1994). *Counting People in the Information Age*. Washington D.C.: National Academy Press.
- EFFRON, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. Philadelphia: Society for Industrial and Applied Mathematics (SIAM).

- FAY, R.E., et THOMPSON, J.H. (1993). The 1990 Post Enumeration Survey Statistical Lessons, in Hindsight. *Proceedings of the 1993 Annual Research Conference*. U.S. Bureau of the Census, 71-91.
- FREEDMAN, D.A., et NAVIDI, W.C. (1992). Aurions-nous dû redresser les chiffres du recensement des E.-U. de 1980? *Techniques d'enquête*, 18, 3-26.
- FREEDMAN, D.A., PISANI, R., et PURVIS, R. (1978). *Statistics*. New York: Norton.
- FREEDMAN, D.A., et WACHTER, K.W. (1993). Heterogeneity and Census Adjustment for the Inter-Censal Base. Technical Report No. 381, Department of Statistics, University of California at Berkeley.
- HAMPPEL, F.R., RONCHETTI, E.M., ROUSSEEUW, P.J., et STAHLE, W.A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. New York: John Wiley and Sons.
- HENGARTNER, N., et SPEED, T.P. (1993). Assessing between-block heterogeneity within the poststrata of the 1990 Post Enumeration Survey. *Journal of the American Statistical Association*, 88, 1119-1125.
- HOGAN, H. (1992). The 1990 Post Enumeration Survey: An overview. *American Statistician*, 46, 261-269.
- HOGAN, H. (1993). The 1990 Post Enumeration Survey: Operations and results. *Journal of the American Statistical Association*, 88, 1047-1057.
- ISAKI, C.T., SCHULTZ, L.K., DIFFENDAL, G.J., et HUANG, E.T. (1988). On estimating census undercount in small areas. *Journal of Official Statistics*, 4, 95-112.
- KIM, J. (1991). 1990 PES Evaluation Project P12: Evaluation of Synthetic Assumption. 1990 Coverage Studies and Evaluation Memorandum Series No. N-4, note de service interne, U.S. Bureau of the Census.
- MULRY, M.H., et SPENCER, B.D. (1993). Accuracy of the 1990 Census and undercount adjustment. *Journal of the American Statistical Association*, 88, 1080-1091.
- SCHAFER, J.L. (1993). Commentaire sur Hengartner, N., et Speed, T.P.: Assessing between-block heterogeneity within the poststrata of the 1990 Post Enumeration Survey. *Journal of the American Statistical Association*, 88, 1125-1127.
- SCHIRM, A.L., et PRESTON, S.H. (1987). Census undercount adjustment and quality of geographic population distributions. *Journal of the American Statistical Association*, 82, 965-978.
- WACHTER, K.W., et FREEDMAN, D.A. (1992). Measuring Local Homogeneity 1990 Census Data. Technical Report, Department of Statistics, University of California at Berkeley.
- WOLTER, K.M., et CAUSEY, B.D. (1991). Evaluation of procedures for improving population estimates for small areas. *Journal of the American Statistical Association*, 86, 278-284.

façon d'agir imposerait une répartition très inefficace de l'échantillon. Jusqu'au recensement de l'an 2000, la recherche devra porter sur l'élaboration d'une combinaison d'un plan d'échantillonnage et de méthodes d'estimation qui produiront des estimations justifiables de la population par Etat.

REMERCIEMENTS

Les auteurs souhaitent remercier les arbitres, un rédacteur associé et le rédacteur en chef pour leurs commentaires, qui ont contribué à améliorer le texte. Cet article présente les résultats généraux de la recherche entreprise par les auteurs. Les opinions exprimées sont celles des auteurs et ne reflètent pas nécessairement la position du Census Bureau ou de la Harvard University. Le travail de Zaslavsky a été appuyé par des subventions dans le cadre des Joint Statistical Agreements 90-23 et 91-31 et par un contrat entre le Census Bureau et le National Opinion Research Center. Le troisième auteur travaillait au Bureau of the Census au moment de la recherche.

ANNEXE

Réalisation, à l'aide de statistiques linéarisées, de tests sur les différences entre les Etats

Une analyse de variance à deux critères de classification portant sur les facteurs de correction dans des parties d'Etat donne un résumé intuitivement significatif des contributions relatives des effets de l'Etat et du GSP sur la variation des facteurs de correction. Comme l'unité d'échantillonnage de l'enquête post-censitaire est la grappe d'îlots pluriel que la partie d'Etat, ces modèles ne donnent pas des tests statistiques valables de la signification des effets de l'Etat.

Considérons une statistique dont l'estimation calculée à partir de l'échantillon pour un Etat ou pour une partie d'Etat est une moyenne pondérée des estimations calculées à partir d'un échantillon dans chaque îlot ou PI qui fait partie de l'Etat. La signification statistique des effets de l'Etat pour cette statistique dans un GSP pourrait être évaluée au moyen d'une analyse de variance à un critère de classification – les parties d'îlots incluses étant utilisées comme unités (correspondant aux UPB) – ou calculée pour l'ensemble des GSP au moyen d'une analyse de variance à deux critères de classification pour les effets de l'Etat et du GSP.

L'estimation du facteur de correction de l'échantillon $(WCE/WE) / (WM/WP)$ est une fonction non linéaire du nombre de parties d'îlots échantillonnées. Dans de petites unités primaires d'échantillonnage (UPB) telle que des parties d'îlots, il se peut que cette non-linéarité soit très remarquable, particulièrement si le nombre d'appartements dans une UPB est très faible ou nul, de telle sorte que l'estimation calculée à partir de l'échantillon du facteur de correction est grande ou infinie. Dans ce cas, si les estimations calculées à partir d'un échantillon d'UPB sont

traitées comme des données, les hypothèses additives de l'analyse de variance ne sont pas respectées. On peut toutefois réaliser des tests utiles en utilisant une version linéarisée de la statistique qui nous intéresse.

Supposons que nous nous intéressions à un paramètre $Z = f(X)$, où X est un vecteur de proportions de la population dans certaines cases. Représentons par x_i, x_j , les proportions correspondantes pour l'échantillon dans tout l'échantillon et dans l'UPB i respectivement, de sorte que $x = \sum N_i x_i / \sum N_i$ est une moyenne pondérée en fonction de la taille des proportions de cases/îlot. Représentons par $f_1(X)$ le gradient de f à X . Alors, au moyen de la linéarisation par série de Taylor, $f(x) - f_1(X) \approx f_1(X)'(x - X) = \sum N_i f_1(X)'x_i / \sum N_i - f_1(X)'X$, c'est-à-dire que nous pouvons traiter le problème comme un problème d'inférence ayant trait aux quantités (pseudo-observations) $z_i = f_1(X)'x_i$. Comme l'influence approximative (linéarisée) de l'UPB i sur l'estimation $f(x)$, c'est-à-dire la différence entre l'estimation obtenue avec ou sans l'UPB i, est $N_i f_1(X)'(x_i - x)$, nous pouvons décrire la méthode utilisée comme une méthode fondée sur les statistiques d'influence (Hampel et coll. 1986) ou sur la méthode du jackknife infinitésimal (Efron 1982, chapitre 6).

un modèle de la superpopulation:
 $E(X_i) = X, \text{ Var}(X_i) = V_i$
et
un modèle d'échantillonnage:
 $E(x_i | X_i) = X_i, \text{ Var}(x_i | X_i) = U_i.$

La covariance d'échantillonnage U_i sera généralement proportionnelle à N_i^{-1} . Une spécification plausible et mathématiquement commode pour V_i est $V_i \propto N_i^{-1}$ (c-à-d. que les UPB plus petites sont plus variables que les plus grandes), de sorte que $\text{Var} z_i = \sigma^2 / N_i$ pour une constante σ^2 . Le poids du modèle linéaire correspondant pour l'UPB i est N_i , de sorte que l'estimation de la moyenne fondée sur un modèle concorde avec l'estimation fondée sur le plan obtenue en regroupant les effectifs de case si N_i est une mesure de taille pondérée.

Dans le cas du facteur de correction $R = (WCE/WE) / (WM/WP)$, les pseudo-observations ont la forme $z_i = f_1(X)'(x_i - x)$ = $R \left(\frac{WCE_i}{WCE} + \frac{WP_i}{WP} - \frac{WE_i}{WE} - \frac{WM_i}{WM} \right)$, où WCE_i, WP_i, WE_i et WM_i sont semblables aux équivalents déjà mentionnés pour la i-ème PI. Nous utilisons l'équation $N_i = (WE_i + WP_i) / 2$ pour obtenir une approximation du poids approprié d'une partie d'îlot.

non urbaines n'a pas donné d'aussi bons résultats que dans les régions plus urbanisées.

Comment pouvons-nous expliquer la différence entre les résultats des deux études? Les deux ensembles de données avaient des tailles d'échantillon très différentes l'une de l'autre, c'est-à-dire que pour les données du recensement on comptait 125,000 grappes d'îlots mais seulement 5,293 pour les données de l'enquête post-censitaire. Il n'est donc pas surprenant que de petites différences entre les Etats, relativement aux variables de remplacement, soient statistiquement significatives bien qu'on ne puisse démontrer l'existence de différences correspondantes relativement aux taux de sous-dénombrement.

De plus, les corrélations entre le taux de sous-dénombrement et les variables de remplacement sont faibles, comme on le voit au tableau 1. Par conséquent, toute généralisation des variables de remplacement aux taux de sous-dénombrement est plutôt conjecturale. Compte tenu de la faible corrélation entre les taux de sous-dénombrement et les variables de remplacement, nous préférons accorder plus de poids à l'analyse des données tirées de l'enquête post-censitaire.

Nous concluons, à partir de ces résultats, qu'il n'y a pas de différences démontrables dans le taux de sous-dénombrement moyen entre les Etats dans chaque division, une fois que l'on a tenu compte des effets du GSP. Bien qu'il existe un faible signe de différence entre le New Jersey et l'Etat de New York, dans la division Mid-Atlantic, il ne faut pas donner trop d'importance à ce résultat dans le contexte du nombre de divisions (9) pour lesquelles le test a été effectué. Nous concluons que si un rajustement des chiffres de population fondé sur l'enquête post-censitaire de 1990 avait été effectué, aucun Etat n'aurait pu démontrer que la stratification a posteriori était manifestement injuste à son endroit en ce sens qu'elle sous-corrigerait les chiffres de cet Etat par rapport à ce que montraient les estimations directes pour l'Etat.

Comme le montre l'analyse présentée à la section 2, il n'y a pas de consensus sur la question de savoir si l'hétérogénéité parmi les Etats dans les taux de sous-dénombrement à l'intérieur des GSP, qui a une certaine importance bien qu'elle ne soit pas suffisamment grande pour être mesurée avec exactitude par l'enquête post-censitaire, aurait ou non une incidence systématique sur le gain d'exactitude obtenu par le rajustement synthétique. Néanmoins, les différences entre les Etats qui ont été observées dans l'analyse de l'enquête post-censitaire, combinées avec la preuve auxiliaire obtenue par les analyses des variables de remplacement, nous amènent à penser qu'il est probable que l'hétérogénéité parmi les Etats sera, à nouveau, un problème lié à la mesure de la couverture pour le recensement de l'an 2000, particulièrement dans les Etats les plus grands, pour lesquels ces différences de couverture peuvent être mesurées avec le plus d'exactitude. Fay et Thompson (1993) soutiennent que pour le recensement de l'an 2000 il faudrait concevoir un échantillon permettant de mesurer la couverture et de faire des estimations directes (plutôt que synthétiques) du sous-dénombrement pour tous les Etats, bien qu'un panel du Comité on National Statistics (CNSSTAT 1994) ait prévenu les spécialistes que, pour certains Etats, cette

Etats. Le fait que les effets estimés aient une valeur importante mais ne soient pourtant pas statistiquement significatifs nous indique que ces tests, effectués pour observer des différences entre les Etats, compte tenu de la taille des échantillons de l'enquête post-censitaire, ne sont pas aussi puissants qu'on pourrait le souhaiter.

Il y a une autre façon de traiter le problème de la puissance, qui est de considérer ce qui se produirait si la taille de l'échantillon du recensement employé pour l'analyse des variables de remplacement était réduite par un facteur de 25, qui est le rapport entre la taille de l'échantillon extrait du recensement et la taille de l'échantillon utilisé pour l'enquête post-censitaire. Si nous divisons par 25 chacun des critères chi carré présentés au tableau 3, alors c'est dans seulement 27 des 99 GSP que les différences entre les Etats auraient été significatives pour le taux de répartition (comparativement à 94 GSP sur 99 pour l'échantillon complet). De même, on aurait observé des différences significatives pour 53 des 99 GSP dans le cas du taux de retour par la poste (comparativement à 92 GSP sur 99 pour l'échantillon complet) et pour 14 taux de substitution sur 84 (comparativement à 74 sur 84). La valeur des taux de substitution se compare à celle des taux de sous-dénombrement; après notre réduction hypothétique de la taille de l'échantillon, nous obtenons des nombres semblables de tests significatifs pour le taux de substitution et le taux de sous-dénombrement. Avec un échantillon beaucoup plus considérable, nous aurions sans doute observé beaucoup plus de différences significatives entre les Etats, bien qu'on ne puisse pas savoir si ces différences auraient été suffisamment grandes pour qu'il faille en tenir compte.

5. DISCUSSION

Dans cet article, nous évaluons l'hétérogénéité du taux de sous-dénombrement et d'autres variables parmi les Etats dans le recensement de 1990.

L'évaluation fait appel à des données du recensement de 1990 et à des données de l'enquête post-censitaire de 1990. Au moment de commencer cette recherche, les données de l'enquête post-censitaire n'étaient pas disponibles et l'on ne s'attendait pas à ce qu'elles le deviennent avant la date prévue pour la fin de cette analyse. Nous avons fait l'essai de variables de remplacement du recensement de 1990 pour vérifier s'il existait une hétérogénéité significative parmi les Etats dans des GSP. Au niveau du GSP, l'effet de l'Etat était significatif ($\alpha = .05$) dans 84 à 95% de ses GSP pour les diverses variables de remplacement.

L'analyse de variance portant sur le sous-dénombrement linéarisé et fondée sur les données de l'enquête post-censitaire au niveau du GSP a montré des effets significatifs de l'Etat ($\alpha = .05$) pour 19 des 99 GSP. Les résultats significatifs étaient concentrés dans les GSP qui se trouvent dans des régions ne faisant pas partie d'une PMSA. On a observé des effets significatifs de l'Etat pour dix des 32 GSP de ce genre. Cela laisse supposer que la stratification effectuée a posteriori dans les régions relativement

Tableau 6
Résumé de l'analyse du sous-dénombrement linéarisé selon le genre d'endroit

Genre d'endroit	Nombre de GSP	Nombre de GSP avec $P < .05$
0	11	3
1	23	1
2	12	1
3	8	1
4	0	0
5	6	2
6	6	1
7	11	3
8	11	4
9	10	3

Les résultats significatifs sont concentrés dans les GSP pour de petites régions (genre d'endroit 7, 8 et 9). Pour 10 des 32 groupes de ce genre, on observe une hétérogénéité significative parmi les États à un seuil de 5%. Cela laisse supposer qu'on peut améliorer la stratification a posteriori dans ces régions.

Le tableau 7 donne la statistique F et la valeur p pour l'effet de l'État pour les modèles État \times GSP, dans un cas avec pondération selon la taille du domaine, dans l'autre sans pondération.

Tableau 7
Effets de l'État selon la division – données pondérées et non pondérées

Division	D.L.	Modèles	
		non pondérés	Modèles pondérés
1	5	.57	.72
2	2	4.64	.01
3	8	.43	.91
4	3	.64	.59
5	3	.66	.58
6	4	.60	.66
7	6	.39	.88
8	7	.62	.74
9	4	.77	.54
		.48	.75

L'effet additif de l'État était significatif dans seulement une division ($p = .01$) dans le modèle État \times GSP non pondéré; quand les données étaient pondérées en fonction de la taille du domaine, la plus petite valeur p pour l'effet de l'État était 0.18. Dans les deux cas, l'effet le plus significatif a été observé dans la division 2, dans laquelle le New Jersey semblait avoir un plus haut taux de sous-dénombrement, si l'on tient compte de l'effet du GSP, que l'État de New York. Il faut remarquer que la région de l'État de New York où l'on observe le plus de sous-dénombrement (la ville de New York) avait sa propre strate

formée a posteriori. Dans huit des dix GSP pour lesquels les États du New Jersey et de New York pouvaient être comparés, y compris les régions non urbaines, le sous-dénombrement estimé était plus élevé pour le New Jersey que pour l'État de New York. Ailleurs, comme les effets de l'État dans différents GSP variaient en importance et parfois en signe, et comme c'est dans seulement une minorité de GSP d'une division quelconque que l'on observait des effets significatifs de l'État, aucune observation significative ne permettait de penser qu'au niveau regroupé la stratification a posteriori était biaisée de façon à défavoriser certains États.

Le tableau 8 donne les estimations ponctuelles des effets de l'État dans des modèles linéaires du taux de sous-dénombrement par partie d'État dans chaque division, avec les effets pour l'État et pour les groupes de strates formés a posteriori. (Les effets sont centrés à 0 par division.) En fait, ce sont des estimations des différences entre les États après qu'on a apporté une correction pour tenir compte des effets expliqués par la composition en GSP des différents États.

Tableau 8
Effets estimés de l'État sur le sous-dénombrement dans une division (en pourcentage)

Division 1			Division 2			Division 3			Division 4			Division 5			Division 6			Division 7			Division 8			Division 9				
CT	-2.42		NJ	4.18		DE	-0.42		AR	1.44		TX	-2.30		IL	0.86		WV	0.11		AZ	2.70		SD	0.60		WA	0.18
ME	.74		NY	-3.91		DC	2.82		LA	-0.71		OK	1.58		IN	1.12		VA	-0.11		CO	0.68		NM	3.35		CA	1.02
MA	-0.48		PA	-0.26		FL	-0.88		TX	-2.30		MT	-1.61		IA	-1.10		SC	0.70		UT	0.08		MT	-1.61		HI	-0.18
NH	-0.14		VT	0.90		GA	-1.43		TX	-2.30		MT	-1.61		KS	-0.50		MD	-1.32		AK	-0.78		ND	-0.07		OR	-0.26
RI	1.43		RI	1.43		NC	0.53		TX	-2.30		MT	-1.61		MS	-0.02		NC	0.53		WY	-2.84		NE	1.76		AK	-0.78

L'ensemble de données utilisé pour ces analyses a été obtenu par fusion de deux ensembles de données pour les 12,124 îlots de l'échantillon de l'enquête post-censitaire, un pour l'échantillon *E* (suivi du recensement), l'autre pour l'échantillon *P* (enquête post-censitaire). Il y avait 12,124 îlots pour la collecte, dont certains ont été subdivisés à des fins de totalisation, ce qui a donné 12,964 îlots pour la totalisation. Chose plus importante, comme certains des îlots les plus petits ont été combinés lors de l'échantillonnage, 5,293 grappes d'îlots ont été échantillonnées. Les dénombrements exacts et les chiffres totaux pour l'échantillon *E* se trouvent dans le fichier de l'échantillon *E*. Les chiffres totaux pour l'échantillon *P* se trouvent dans le fichier de l'échantillon *P* qui comprend aussi les nombres de correspondances (les cases de l'échantillon qui sont inclus dans le recensement).

4.1 Variance expliquée par l'Etat et par le GSP

Pour chaque division, nous avons ajusté une analyse de variance à deux critères de classification aux taux de sous-dénombrement pour des parties d'Etat. Le tableau 4 montre le rapport entre la somme des carrés attribuable aux GSP et la somme attribuable aux Etats dans une division.

Tableau 4

Variance du taux de sous-dénombrement expliquée par l'Etat et le GSP

Div.	Nombre de groupes	Nombre d'Etats*	SC (Groupe)	SC (Etat)	CM (Groupe)	CM (Etat)
1	5	6	4.51	5.64	5.64	8.89
2	12	3	4.88	6.77	12.69	8.73
3	16	9	12.69	3.74	8.17	2.72
4	8	4	8.73	2.19	7.67	2.09
5	10	4	8.17	2.78	1.53	7
6	15	5	7.67	40.28	10.07	17
7	9	7	2.78	4.51	5.64	8
8	7	8	1.53	5.64	8.89	6
9	17	5	40.28	10.07	10.07	19

* Les Etats incluent le District of Columbia.

Le rapport est toujours supérieur à un et, dans la division 9, il est de 40,28, ce qui montre des effets beaucoup plus considérables pour le GSP que pour l'Etat. Le carré moyen pour le groupe dépasse aussi le carré moyen pour l'Etat dans toutes les divisions sauf la division 2. Cela justifie la décision d'utiliser la SP plutôt que l'Etat comme case pour l'estimation et la correction du sous-dénombrement.

4.2 Tests pour les effets de l'Etat sur les taux de sous-dénombrement

Si l'on suppose que le taux de substitution (la fraction des unités pour lesquelles il y a eu imputation pour corriger la non-réponse) est négligeable, le facteur de correction (*R*) pour un domaine est

$$R = \frac{WCE/WE}{WM/WP}$$

où *WCE*, *WP*, *WE* et *WM* sont les contributions de la *i*-ième PI aux totaux mentionnés plus haut. On a ajusté un modèle linéaire aux statistiques d'influence de la PI pour réaliser un test sur les effets de l'Etat. Selon l'hypothèse nulle, toutes les parties d'un Etat dans un GSP ont le même taux de sous-dénombrement et la moyenne théorique de la statistique d'influence pour chaque Etat est de 0 dans chaque GSP. La statistique d'influence peut être étudiée au moyen d'une analyse de variance à un critère de classification dans un seul GSP ou d'une analyse de variance à deux critères de classification pour tous les GSP dans une division. Le tableau 5 résume les tests pour les effets de l'Etat sur des statistiques linéarisées dans chaque GSP.

Tableau 5

Division	Nombre de GSP	Nombre de GSP avec $P < .05$
1	5	0
2	12	3
3	16	4
4	8	5
5	10	2
6	15	1
7	9	0
8	7	1
9	17	3
Somme	99	19

Les tests permettent de constater l'existence d'une hétérogénéité significative parmi les Etats dans 19 des 99 groupes à un seuil de signification de 5%. La valeur de l'effet de l'Etat estimé va de quelques points de pourcentage jusqu'à 20%, mais les erreurs-types de ces estimations sont très grandes. Le tableau 6 résume les résultats de ces analyses selon le genre d'endroit. Les genres d'endroit 0, 1, 2 et 3 correspondent à de grandes villes centrales dans une Primary Metropolitan Statistical Area (PMSA), les genres d'endroit 4, 5 et 6 sont des villes non centrales dans des PMSA renfermant de grandes villes centrales et les genres d'endroit 7, 8 et 9 correspondent aux autres régions.

et le taux de sous-dénombrement est

$$1 - 1/R,$$

où *WE* et *WP* sont les tailles estimées pour la population, pondérées à partir de l'échantillon *E* et de l'échantillon *P*, respectivement. *WCE* est le nombre pondéré de dénombrements exacts et *WM*, le nombre pondéré d'appariements dans l'enquête post-censitaire. La statistique d'influence (voir l'annexe) de la *i*-ième PI sur le facteur de correction ou sur le taux de sous-dénombrement est

$$I_i = R \left(\frac{WCE_i}{WP_i} + \frac{WP_i}{WE_i} - \frac{WE_i}{WM_i} \right),$$

dénombrement est

$$D_{ij} = \frac{\sum_{k=1}^{K_{ij}} n_{ijk}(\hat{p}_{ijk} - \hat{p}_{ij})^2}{K_{ij} \hat{p}_{ij}(1 - \hat{p}_{ij})},$$

où \hat{p}_{ijk} est le taux pour le i -ième GSP, le j -ième Etat et la k -ième PI combinée; n_{ijk} est la taille de la PI combinée dans le i -ième GSP et le j -ième Etat; et \hat{p}_{ij} est le taux estimé pour le i -ième GSP et le j -ième Etat. La fraction est le rapport entre la variance inter-îlots observée et la variance prévue pour un échantillonnage binomial.

L'effet du plan estimé D_{ij} est une mesure de l'hétérogénéité à l'intérieur des Etats à l'intérieur des GSP. Plus il y a d'hétérogénéité à l'intérieur des Etats, plus la variance d'échantillonnage du taux au niveau de l'Etat est élevée et plus il est difficile de détecter une différence significative. L'importance de l'effet du plan a donc une incidence sur la puissance des tests d'hypothèses.

L'effet du plan calculé n'est qu'une approximation de la correction requise. Premièrement, pour calculer D_{ij} , on obtient la somme pour les PI combinées plutôt que pour les PI particulières. Deuxièmement, l'échantillon est un échantillon en grappes stratifié et la majorité ou la totalité des strates formées a posteriori chevauchent plusieurs strates d'échantillonnage. C'est seulement pour un échantillon non stratifié que la formule est strictement exacte. Troisièmement, pour calculer le véritable effet du plan, il faut utiliser aussi bien les termes qui ne sont pas sur la diagonale (covariance) que ceux qui s'y trouvent (variance); or les termes qui ne sont pas sur la diagonale sont omis. Pour tenir compte de cette situation, les effets du plan calculés ont été multipliés par le facteur 1.25, que nous avons obtenu par choix raisonné.

Un effet du plan a été calculé pour chaque variable de remplacement et pour chaque GSP. Cet effet est faible (sa valeur s'établissant autour de 2) dans la plupart des GSP pour le taux de répartition et le taux de substitution. Il est un peu plus élevé pour le taux de retour par la poste, mais il tend à être important (sa valeur pouvant atteindre 20) pour le taux des immeubles à logements multiples et pour le taux des questionnaires expédiés par la poste, car ces caractéristiques sont habituellement assez uniformes dans un îlot mais varient beaucoup entre les îlots.

Le tableau 2 résume les tests, corrigés pour tenir compte de l'effet du plan, pour les effets de l'Etat dans les GSP.

À l'échelle nationale, pour chaque variable de remplacement, l'effet de l'Etat est significatif pour au moins 84% des GSP. (Le nombre total de GSP varie parce que, lorsqu'un GSP fait partie d'un seul Etat ou qu'un seul Etat compte des observations non nulles pour une variable particulière, il est impossible d'ajuster le modèle correspondant). Les résultats sont résumés au niveau de la division. (Les divisions 1 à 9 sont les suivantes: New England, Mid-Atlantic, South Atlantic, East South Central, West South Central, East North Central, West North Central, Mountain et Pacifique).

4. ANALYSE DU TAUX DE SOUS-DÉNOMBREMENT

Les résultats décrits plus haut pour les variables de remplacement ont été obtenus au début du processus de réalisation du recensement, mais leur pertinence est limitée pour ce qui est de l'homogénéité du sous-dénombrement lui-même. Après le traitement des données de l'enquête post-censitaire, on a pu effectuer une analyse directe de la distribution du sous-dénombrement.

Dans les modèles au niveau de la division avec effet de l'Etat et effet du GSP, les effets tant de l'Etat que du GSP étaient significatifs à un seuil de 1% dans toutes les divisions et pour toutes les variables (sauf pour le taux des questionnaires expédiés par la poste dans deux divisions où il a été impossible de calculer le critère utilisé dans le test).

Minimum	25 ^e percentile	50 ^e percentile	75 ^e percentile	Maximum
4.3	27.5	68.9	140.3	945.2
0.28	102.83	254.49	644.05	8,779.88
5.46	49.80	97.35	260.88	1,815.12

Taux de substitution

Taux de retour par la poste

Taux de répartition

Tableau 3
Valeurs des effets de l'Etat par rapport aux critères utilisés dans les tests

Le tableau 3 donne les valeurs des effets de l'Etat, exprimés comme valeurs des variables χ^2 des critères utilisés dans les tests corrigés pour tenir compte de l'effet du plan, dans le cas de trois variables pour lesquelles la corrélation avec le taux de sous-dénombrement est relativement élevée. Dans ce tableau, les valeurs des variables χ^2 ont de 1 à 8 degrés de liberté.

Div.	Nombre de groupes	Répartition par la poste	Immeubles à logements multiples	Questionnaires expédiés par la poste	Somme
1	5	5	5	1(1)	3(4)
2	12	11	12	7(10)	12
3	16	15	16	3(3)	12(12)
4	8	8	7	5(6)	5(8)
5	10	10	9	4(4)	7(8)
6	15	13	15	5(7)	15
7	9	9	9	4(4)	8(8)
8	7	7	7	2(3)	6(6)
9	17	15	14	5(5)	6(12)
99	94	92	95	36(43)	74(84)

Questionnaires expédiés par la poste

Immeubles à logements multiples

Répartition par la poste

Nombre de groupes

Div.

Tableau 2
Nombre de GSP avec effet significatif ($\alpha = .05$) de l'Etat (régression logistique)

Suite à l'application d'un test élémentaire qui traite les GSP comme s'ils étaient indépendants, chaque corrélation est significative sauf celle pour le taux des questionnaires expédiés par la poste, mais la valeur de chaque corrélation n'est pas élevée. De plus, dans une certaine mesure, ces variables décrivent des conditions qui tendent à mener à des taux d'omission plus élevés (répartitions attribuables à un faible taux de questionnaires remplis, substitutions attribuables à la difficulté de réaliser des interviews) ou à des taux d'omission plus faibles (taux élevés de retour par la poste). Par contre, des conditions difficiles de réalisation du recensement peuvent aussi mener à des dénombrements erronés, de sorte que ces effets sur le sous-dénombrement net ne sont pas très bien définis. Si nous n'analysons pas ces variables, c'est tout simplement parce que nous croyons qu'elles sont distribuées exactement de la même façon que le sous-dénombrement. Nous espérons plutôt qu'en obtenant des résultats sur la distribution d'une gamme différente de variables du recensement, nous pourrions avoir une meilleure idée de la distribution du sous-dénombrement.

Pour l'analyse des variables de remplacement, nous avons extrait un échantillon en grappes stratifié de données du recensement de 1990. Cet échantillon est composé de 204,394 îlots correspondant à 125,000 grappes d'îlots. Chaque partie d'îlot contenant moins de dix personnes a été combinée avec les parties d'îlots suivantes (selon l'ordre du numéro d'îlot) jusqu'à ce qu'on obtienne une combinaison de parties d'îlot renfermant au moins dix personnes. Nous avons procédé de la sorte afin d'obtenir des taux relativement stables pour les variables de remplacement qui nous permettent d'analyser les taux eux-mêmes.

Les variables de remplacement sont analysées par régression logistique. Deux formes de modèle de régression logistique ont été utilisées. Pour l'analyse à l'intérieur des

où P_{ij} est le taux pour une variable de remplacement dans le i -ième GSP et le j -ième État, A est l'ordonnée à l'origine, B_i est l'effet du i -ième GSP et C_j est l'effet du j -ième État. Pour les modèles, nous n'avons utilisé que les 99 GSP chevauchant les frontières d'au moins deux États. Des modèles ont été produits pour les variables de remplacement dans les 99 GSP et dans chacune des neuf divisions. Nous avons utilisé PROC CATMOD du SAS pour estimer les paramètres au moyen de la méthode du maximum de vraisemblance et pour obtenir des statistiques de Wald afin de tester la signification des effets de l'État.

Les données ont été recueillies au moyen d'un échantillon par grappes plutôt que d'un échantillon aléatoire simple, de sorte que les critères utilisés dans le test doivent être divisés par un effet du plan. Nous estimons un effet du plan,

$$\log [P_{ij} / (1 - P_{ij})] = A + B_i + C_j,$$

et, pour l'analyse à l'intérieur des divisions,

$$\log [P_{ij} / (1 - P_{ij})] = A + C_j$$

GSP, le modèle pour le GSP i est

3. ANALYSE DES VARIABLES DE REMPLACEMENT

présenté des contre-exemples de distributions possibles du sous-dénombrement pour lesquelles le rajustement à l'aide de l'estimation synthétique rendrait la distribution de la population moins exacte.

Fay et Thompson (1993) ont simulé les effets de l'hétérogénéité sur l'exactitude d'estimations synthétiques, à l'aide de huit variables de remplacement (dont les cinq utilisées dans la présente étude) et de l'ensemble de données analysé dans la section 3. Ils ont effectué une analyse par fonction de perte comme dans Mulry et Spencer (1993) pour comparer l'exactitude de chiffres simulés non rajustés à celle de chiffres rajustés synthétiquement. Ils ont constaté que si l'on ne tenait pas compte de l'hétérogénéité on sous-estimait le gain d'exactitude attribuable au rajustement synthétique pour cinq des huit variables et on le surestimait pour une variable (le taux de chômage), alors qu'il y avait peu de différence pour deux autres variables (le taux de pauvreté et le taux de migration).

Au cours de l'analyse des données du recensement, nous avons choisi des variables qui étaient disponibles pour tout le recensement et qui, comme le sous-dénombrement, décrivaient le processus du recensement ou y étaient liées. Les variables de remplacement choisies sont le taux de répartition, le taux de retour par la poste, le taux des immeubles à logements multiples, le taux des questionnaires expédiés par la poste (fraction des unités qui reçoivent un questionnaire par la poste) et le taux de substitution. Le taux de répartition est la fraction des ménages pour lesquels des imputations ont été faites de manière à corriger la non-réponse partielle et le taux de substitution est la fraction des ménages pour lesquels toutes les valeurs ont été imputées parce qu'on a déterminé qu'une unité était occupée mais qu'aucune interview n'avait pu être réalisée. Le tableau 1 montre les corrélations entre chacune de ces variables et le taux de sous-dénombrement par GSP. Ces corrélations "écologiques" (Freedman, Pisani et Purvis 1978, pp. 141-142) de moyennes des GSP diffèrent de celles qui pouvaient être calculées à partir des données au niveau des îlots. Ces dernières sont moins importantes, peut-être à cause du bruit introduit par la variabilité aléatoire dans les petites populations de chaque îlot.

Tableau 1

Coefficients de corrélation entre la variable de remplacement et le taux de sous-dénombrement par GSP	
Variable	Corrélation
Taux de répartition	.44
Taux de retour par la poste	-.57
Taux d'immeubles à logements multiples	.39
Taux de questionnaires expédiés par la poste	.08
Taux de substitution	.47

2. REVUE DE LA LITTÉRATURE
SPECIALISÉE

Dans les sources qui traitent de l'hétérogénéité, on s'est

intéressé à deux questions clés:

- 1. La question empirique: quelle est l'importance de l'hétérogénéité et comment peut-on la décrire.

- 2. La question théorique et le point de vue des politiques: quelles sont les implications de l'hétérogénéité pour l'exactitude des rajustements synthétiques et la validité des évaluations de ces rajustements?

On peut détecter et analyser l'hétérogénéité à plusieurs niveaux d'agrégation. La parfaite homogénéité des taux de sous-dénombrement pour de très petits domaines est numériquement impossible, à cause du caractère discontinu de la population réelle et des chiffres du recensement. En fait, comme les erreurs dans le recensement (omissions ou dénombrements erronés) sont en général ou bien indépendantes les unes des autres, ou bien associées positivement (comme lorsqu'un ménage de plusieurs membres est omis ou qu'une caractéristique locale a une incidence sur tout un îlot), nous nous attendrions à constater au moins une variabilité binomiale dans les taux de sous-dénombrement observés.

Hengartner et Speed (1993) ont analysé les données de l'enquête post-censitaire de 1990 pour deux endroits en ajustant des modèles dans lesquels les variables explicatives étaient l'îlot et le "démotide" (unité définie par les variables de stratification à posteriori non géographiques telles que l'origine ethnique, le sexe, l'âge et le mode d'occupation du logement). Ils ont constaté que la proportion de la variance expliquée par l'îlot était légèrement supérieure à la proportion expliquée par le démotide; le nombre d'îlots n'était pas beaucoup plus élevé que le nombre de démotides dans leur ensemble de données. En réponse, Schafer (1993) a soutenu qu'il ne serait pas pratique d'employer un plan d'estimation utilisant des effets d'îlot parce qu'il faudrait alors recueillir des données dans chaque îlot.

L'hétérogénéité du sous-dénombrement à un niveau d'agrégation quelconque peut être définie comme l'excédant de variabilité des taux de sous-dénombrement observés à ce niveau par rapport à ce qu'on s'attendrait d'observer comme conséquence de la variabilité à un niveau d'agrégation inférieur. Par exemple, en limitant notre attention à une seule strate formée à posteriori, un ensemble d'îlots est hétérogène si les taux de sous-dénombrement dans cette strate à posteriori diffèrent plus que ce que l'on prévoirait si les ménages, y compris ceux qui, au recensement, ont été dénombrés, partiellement dénombrés ou omis, avaient été répartis au hasard entre les îlots. De même, un groupe d'États est hétérogène (en neutralisant de la même façon la strate à posteriori) lorsque ces États diffèrent plus qu'on ne le prévoirait si les îlots, y compris ceux qui ont des taux de sous-dénombrement plus élevés et plus faibles, avaient été répartis au hasard entre les États. Plusieurs études ont tenté de mesurer l'hétérogénéité dans les taux de sous-dénombrement et dans d'autres variables du recensement. Wachter et Freedman (1992) ont analysé un gros échantillon de données du recensement (semblable à celui étudié dans

la section 3). Ils ont estimé l'excédant de variabilité entre des "super-îlots" sur celle qui était prévue par un modèle binomial avec des taux uniformes, pour quatre variables de "population artificielles" (taux de logement dans les immeubles à logements multiples, taux de non-retour par la poste, répartitions et substitutions) décrites dans la section 3. Comparativement à la plus grande hétérogénéité possible (si chaque îlot était homogène), l'excédant de variabilité" allait d'environ 20% (pour les immeubles à logements multiples) à 2% (pour les substitutions). Une autre étude, de Freedman et Wachter (1993), examinait l'hétérogénéité parmi les États à l'aide de "populations artificielles" fondées sur les mêmes variables et sur deux autres, et l'on a observé une variabilité importante. Alho, Mulry, Wurdeman et Kim (1993) ont utilisé des modèles de régression logistique conditionnels pour décrire l'hétérogénéité associée à des covariables mesurées qui n'étaient pas prises en compte lors de la stratification à posteriori. Ils cherchaient plus à réduire le biais d'estimations de système dual de la population qu'à obtenir des estimations plus exactes pour de petites régions.

Un sujet controversé lors de l'évaluation du redressement proposé des chiffres du recensement de 1990 était l'effet de l'hétérogénéité sur l'exactitude des chiffres de population redressés obtenus à l'aide de l'estimation synthétique et, particulièrement, sur les comparaisons de l'exactitude des formes dans une strate formée à posteriori, les mesures groupées de l'exactitude d'un recensement redressé sous-estimant l'erreur de façon systématique. Toutefois, comme la non-uniformité de la couverture a aussi une incidence sur l'exactitude des chiffres non redressés d'un recensement, les implications de cette conclusion pour l'opportunité du rajustement ne sont pas évidentes.

Dans une précédente étude sur les "variables de remplacement", Isaki, Schultiz, Diffendal et Huang (1988) ont simulé le comportement d'estimateurs synthétiques sur des "populations artificielles" qui étaient des transformations du taux de substitution (imputation unitaire). Ils ont constaté qu'un estimateur synthétique donnait généralement de meilleurs résultats que les chiffres "non redressés".

Schirm et Preston (1987) ont soutenu, en se fondant sur les résultats de calculs analytiques et de simulations, que l'estimation synthétique rend les estimations pour de petites régions plus exactes dans des conditions plausibles, même si l'hypothèse synthétique n'est pas vérifiée. Wolter et Caussey (1991) ont étudié le rendement des estimateurs synthétiques et d'un seul rajustement du ratio quand les taux de sous-dénombrement sont estimés avec erreur, à l'aide de taux du sous-dénombrement provenant du Post-Enumeration Program (PEP) de 1980 et en simulant divers niveaux d'erreur d'échantillonnage; ils ont estimé des coefficients de variation "de point d'équilibre" auxquels l'erreur d'échantillonnage dans les proportions ou chiffres rajustés les rendraient moins exacts que des proportions ou des chiffres non rajustés. Les conclusions de ces articles ont été critiquées par Freedman et Navidi (1992), qui ont

Hétérogénéité inter-états du taux de sous-dénombrement et les variables de remplacement dans le recensement des États-Unis de 1990

JAY JONG-IL KIM, ALAN ZASLAVSKY et ROBERT BLODGETT¹

RÉSUMÉ

Dans le cadre de la décision relative au redressement des chiffres du recensement décennal de 1990, le U.S. Census Bureau a étudié l'hétérogénéité possible des taux de sous-dénombrement parmi des parties de différents États se trouvant dans la même case de rajustement ou strate formée à posteriori. Cinq "variables de remplacement" que l'on croyait associées au sous-dénombrement ont été analysées à l'aide d'une grande partie des données du recensement, et l'on a constaté une hétérogénéité significative. L'analyse de l'enquête post-censitaire sur les taux de sous-dénombrement a montré que les variables de stratification à posteriori expliquaient une plus grande partie de la variance que l'État, d'où le choix de la strate à posteriori comme case de rajustement. On a observé une hétérogénéité significative parmi les États dans 19 des 99 groupes de strates à posteriori (surtout dans les régions non urbaines), mais, après avoir regroupé des strates à posteriori, on n'a pratiquement rien observé qui donne à penser que l'estimateur de stratification à posteriori était biaisé de façon à défavoriser certains États. Néanmoins, de futures études sur l'évaluation de la couverture devraient permettre de résoudre cette question.

MOTS CLÉS: Stratification à posteriori; statistiques d'influence; linéarisation; estimation synthétique.

1. INTRODUCTION

L'enquête post-censitaire (BPC) du recensement décennal de 1990 des États-Unis était conçue pour produire des estimations de couverture pour 1,392 strates formées à posteriori. Le pays a d'abord été divisé en 16 domaines, appelés groupes de strates à posteriori (GSP) selon la géographie, l'origine ethnique/espagnole et le mode d'occupation (pro-priétaire par opposition à locataire). Avec seulement quatre exceptions, tous les GSP sont définis dans une division de recensement, les divisions, au nombre de neuf, étant des régions géographiques composées chacune de plusieurs États contigus. Chaque GSP a ensuite été subdivisé en douze groupes définis selon l'âge et le sexe, qui sont les strates formées à posteriori. Par exemple, presque tous les locaux de la ville de New York forment un GSP et toutes les filles âgées de 0 à 9 ans dans ce GSP forment une strate à posteriori (SP). D'autres détails sur l'enquête post-censitaire sont présentés dans Hogan (1992, 1993). Les taux de sous-dénombrement dans les petites régions ont été calculés par estimation synthétique; le même facteur de correction a été appliqué aux personnes d'une SP donnée dans toutes les régions. Cette procédure est juste selon l'"hypothèse synthétique" de l'homogénéité du taux de sous-dénombrement dans une SP. On a beaucoup discuté de la validité de l'hypothèse synthétique (section 2). Dans cet article, nous faisons état de la recherche effectuée dans le cadre d'un projet d'évaluation de l'enquête post-censitaire (le "projet P12"), qui étudiait l'hétérogénéité à l'intérieur des strates à posteriori. Cette recherche portait

en particulier sur la possibilité de constater des différences de couverture entre des parties d'une strate à posteriori qui se trouvent dans des États différents. Suivant l'hypothèse de l'homogénéité, les taux sont identiques dans un SP, quel que soit l'État. Donc, on peut effectuer un test portant sur cette hypothèse en comparant les taux d'un État à l'autre dans une même SP; ce test fait porter l'attention sur la question de savoir si l'estimation synthétique est "juste" pour certains États. L'unité d'analyse est l'intersection d'un îlot de recensement et d'une SP ou d'un GSP, appelée partie d'îlot (PI) pour l'analyse des données sur le taux de sous-dénombrement. Un îlot de recensement est une petite région délimitée par des caractéristiques physiques comme des rues, des cours d'eau, etc. et (ou) par des limites territoriales. Dans les régions urbaines, un îlot de recensement correspond approximativement à un pâté de maison ou îlot. En fait, la plus grande partie de nos analyses porte sur des GSP, puisque la répartition selon l'âge et le sexe dans les GSP ne varierait pas beaucoup d'un État à l'autre. Par conséquent, l'analyse est conçue pour déterminer si, dans un GSP, les PI diffèrent entre les États. Deux analyses distinctes ont été effectuées. On a étudié la distribution de cinq "variables de remplacement" (section 3), à l'aide d'un gros ensemble (4,26%) de données du recensement. On a étudié la distribution du sous-dénombrement à l'aide de l'ensemble de données, beaucoup plus petit, de l'enquête post-censitaire (section 4). Pour des tableaux et une documentation plus détaillés sur le projet, voir Kim (1991).

¹ Jay Jong-Il Kim, Statistical Research Division, U.S. Bureau of the Census, Suitland, MD 20233, U.S.A.; Alan Zaslavsky, Department of Statistics, Harvard University, Cambridge, MD 02138, U.S.A.; et Robert Blodgett, U.S. Food and Drug Administration, 200 C St., S.W., Washington, DC 20204, U.S.A.

- LANGF, N., et RYAN, L. (1989). Assessing normality in random effects models. *The Annals of Statistics*, 17, 624-642.
- MARITZ, J.S., et LWIN, T. (1989). *Empirical Bayes Methods* (2^{ème édition}). London: Chapman and Hall.
- MICHALOWSKI, M. (1993). Revised postcensal and intercensal estimates: Canada, provinces and territories, 1971 - 1991. Rapport interne, Section des estimations démographiques, Statistique Canada.
- PFEFFERMANN, D., et BURCK, L. (1990). Estimation robuste pour petits domaines par la combinaison de données chronologiques et transversales. *Techniques d'enquête*, 16, 229-249.
- PRASAD, N.G.N., et RAO, J.N.K. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, 85, 163-171.
- RAO, C.R. (1973). *Linear Statistical Inference and its Applications*. New York: John Wiley and Sons.
- ROBINSON, G.K. (1991). That BLUP is a good thing: the estimation of random effects (avec discussion). *Statistical Sciences*, 6, 15-51.
- ROYCE, D. (1992). Une comparaison d'estimateurs d'un ensemble de totaux de population. *Techniques d'enquête*, 18, 121-138.
- STATISTIQUE CANADA (1993). *Rapports techniques du recensement de 1991: couverture*. N° 92-341F au catalogue. Statistique Canada.
- WONG, G.Y., et MASON, W.M. (1985). The hierarchical logistic regression model for multilevel analysis. *Journal of the American Statistical Association*, 80, 513-524.

5. RÉSUMÉ ET CONCLUSIONS

Nous avons adopté la méthode empirique de Bayes parce qu'elle conserve les estimations les plus fiables (c.-à-d. celles qui se rapportent aux grands domaines) tout en permettant la substitution d'une estimation basée sur un modèle lorsque l'estimation directe correspondante n'est pas fiable. Cela est conforme à la pratique courante en techniques d'enquête, qui est d'utiliser le plus possible les estimations directes. La méthode itérative du quotient utilisée pour corriger les estimations tirées du modèle empirique de Bayes a servi à faire concorder les estimations de modèle avec les estimations d'enquête que l'on savait fiables.

Quant au modèle explicite utilisé pour décrire les facteurs de redressement vrais, notons qu'il est purement descriptif. Sa fonction première est de décrire la variation des facteurs de redressement à l'aide de variables explicatives en tenant compte de l'erreur d'échantillonnage rattachée à chaque facteur de redressement. Il ne serait donc pas prudent de tirer des conclusions d'une grande portée sur la nature du sous-dénombrement en se fondant sur l'ensemble final des paramètres du modèle.

La plus grande faiblesse de la méthode décrite dans cet article a trait aux deux variances qui sont estimées. L'hypothèse selon laquelle les erreurs du modèle de régression sont distribuées approximativement selon une loi normale est difficile à évaluer. En l'absence de tout renseignement concret sur la distribution vraie, les hypothèses concernant la variance de modèle ne seront pour ainsi dire pas vérifiables. La variance de modèle proposée semble raisonnable et les tests de diagnostic n'ont révélé aucune difficulté majeure.

Le modèle de la variance d'échantillonnage pose plus de problèmes. Toutes les méthodes empiriques de Bayes supposent cette variance connue alors qu'en réalité il faut l'estimer. Les efforts qui avaient pour but d'inclure ce paramètre estimé dans le calcul de l'EQM de Prasad-Rao n'ont donné aucun résultat nouveau jusqu'à maintenant. Dans les recherches futures, on s'intéressera plus particulièrement aux problèmes qui se rattachent à l'estimation des variances d'échantillonnage. Il faut pousser plus loin la recherche sur le calcul de l'EQM de Prasad-Rao. En outre, on continuera d'étudier la possibilité de se servir des microdonnées provenant des études relatives à la couverture et la possibilité d'estimer directement les taux de sous-dénombrement par la régression logistique, comme dans Wong et Mason (1985).

Un autre projet de recherche serait d'analyser les conséquences d'une reformulation du modèle empirique de Bayes en termes de modèle d'espace d'états (Robinson 1991). Pfeffermann et Burck (1990) ont proposé une méthode de calcul de l'EQM pour une série temporelle qui est définie à l'intérieur d'un modèle d'espace d'états qui doit être en conformité avec certaines données répétées périodiques. La formulation en modèle d'espace d'états serait aussi utile pour l'intégration explicite des méthodes démographiques.

BIBLIOGRAPHIE

- L'auteur tient à remercier D. Binder, M. Hidiroglou, R. Carter, M. Armstrong, J.N.K. Rao et en particulier D. Royce pour les commentaires que ces personnes lui ont adressés sur une version antérieure de cet article. L'auteur remercie également le rédacteur associé et les deux rapporteurs qui, par leurs commentaires, ont contribué à améliorer la version finale de l'article.
- REMERCIEMENTS**
- BURGESS, R.D. (1988). Evaluation des estimations du sous-dénombrement obtenues par la contre-vérification des dossiers du recensement du Canada. *Techniques d'enquête*, 14, 147-167.
- CHOI, C.Y., STEEL, D.G., et SKINNER, T.J. (1988). Redressement des chiffres du recensement de 1986 en Australie pour le sous-dénombrement. *Techniques d'enquête*, 14, 187-204.
- CRESSIE, N. (1992). Estimation du maximum de vraisemblance avec contrainte (MVC) dans le lissage des taux de sous-dénombrement du recensement selon l'approche empirique de Bayes. *Techniques d'enquête*, 18, 83-103.
- DATTA, G.S., GHOSH, M., HUANG, E.T., ISAKI, C.T., SCHULTZ, L.K., et TSAY, J.H. (1992). Méthode hiérarchique de Bayes et méthode empirique de Bayes pour le redressement du sous-dénombrement: données de la "répétition générale" du recensement, effectuée en 1988 au Missouri. *Techniques d'enquête*, 18, 105-119.
- DICK, J.P. (1993). Procedures used in modelling net under-coverage in the 1991 Census. Note de service interne de Statistique Canada.
- DRAPER, N.R., et SMITH, H. (1966). *Applied Regression Analysis*. New York: John Wiley and Sons.
- ERICKSEN, E.P., et KADANE, J.B. (1985). Estimating the population in a census year (avec discussion). *Journal of the American Statistical Association*, 84, 927-943.
- FAY, R.E., et HERRIOT, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 82, 269-277.
- FREEDMAN, D., et NAVIDI, W. (1992). Aurions-nous dû redresser les chiffres du recensement des E.-U. de 1980? (avec discussion). *Techniques d'enquête*, 18, 3-26.
- GERMAIN, M.-F., et JULIEN, C. (1993). Results of the 1991 Census coverage error measurement program. *Proceedings of Seventh Annual Research Conference*. United States Bureau of the Census, 55-70.
- GHOSH, M., et RAO, J.N.K. (1994). Small area estimation: An appraisal (avec discussion). *Statistical Science*, 9, 55-93.
- HARVILLE, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72, 320-337.
- HOGAN, H. (1992). The 1990 Post-Enumeration Survey: an overview. *The American Statistician*, 46, 261-269.
- JUDGE, G.G., GRIFFITHS, W.E., CARTER HILL, R., et LEE, T.-C. (1984). *The Theory and Practice of Econometrics*. New York: John Wiley and Sons.

Cette procédure convergera vers une solution unique. Comme il s'agit essentiellement d'un modèle log-linéaire, l'hypothèse sous-jacente est que la relation que détermine le modèle empirique de Bayes en ce qui concerne l'interaction entre la province et le groupe d'âge-sexe est valide et sera préservée.

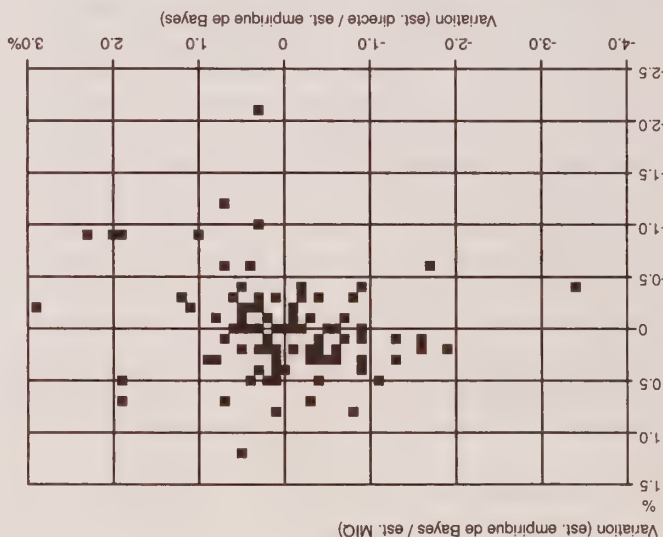


Figure 3. Variation en pourcentage des estimations des facteurs de redressement.

Il y a toutefois un aspect négatif; en effet, l'EQM des estimations MIQ des facteurs de redressement est très difficile à estimer. En raison du caractère non linéaire de la méthode itérative du quotient, il est impossible de calculer directement l'EQM. On peut toujours recourir à un développement en série de Taylor, sauf que cette solution suppose un gros effectif d'échantillon pour chaque domaine, et nous savons très bien que certains domaines ont un effectif d'échantillon très petit. Une autre façon de procéder consiste à redresser l'EQM estimée des estimations empiriques de Bayes et de multiplier ces dernières par le carré du rapport entre l'estimation empirique de Bayes

et les estimations empiriques de Bayes correspondantes. De façon générale, la méthode itérative du quotient a pour effet de rapprocher l'estimation empirique de Bayes de l'estimation d'enquête. C'est ce que montre la figure 3. Celle-ci illustre graphiquement deux types de variation en pourcentage des estimations des facteurs de redressement. Sur l'axe des X est représenté l'écart relatif entre l'estimation directe d'enquête et l'estimation empirique de Bayes, tandis que sur l'axe des Y est représenté l'écart relatif entre l'estimation empirique de Bayes et l'estimation MIQ finale. Le graphique montre que les deux variables sont corrélées négativement; par conséquent, la méthode itérative du quotient tend à rapprocher l'estimation empirique de Bayes de l'estimation d'enquête.

4.2 Estimations pour domaine détaillées

Le Programme des estimations démographiques exige un niveau de détail encore plus poussé que celui des estimations produites par les divers modèles étudiés ci-dessus. De fait, le Programme requiert, pour chaque province, des estimations par âge, sexe et division de recensement. Comme la méthode empirique de Bayes est limitée quelque peu par les données d'enquête (il faut en effet une estimation qui ait une erreur type non nulle pour chaque domaine), on doit recourir à des méthodes synthétiques pour produire des estimations plus détaillées.

En ce qui concerne le Programme des estimations démographiques, des estimations ont été produites, par province et par sexe, pour neuf groupes d'âge au lieu des quatre groupes utilisés dans le modèle empirique de Bayes. Pour la première étape de l'estimation, on a posé un modèle synthétique simple qui utilise comme valeurs initiales les estimations empiriques de Bayes soumises à la méthode itérative du quotient. Pour produire les estimations plus détaillées, on a réparti l'estimation empirique de Bayes soumise à la MIQ entre tous les sous-groupes d'âge, par province et par sexe, au prorata de l'effectif recensé. Posons l'estimation finale de la méthode itérative du quotient pour la province p et le groupe d'âge-sexe a comme $M_{pa}^{2a+2} = M_{pa}^a$. De plus, si le groupe d'âge-sexe a se compose de Q sous-groupes d'âge disjoints, alors le nombre estimé de personnes oubliées pour la province p et le sous-groupe d'âge q du groupe d'âge-sexe

$$M_{paq}^{pa} = M_{pa}^a \left(\frac{C_{paq}^{pa}}{C_{pa}^{pa}} \right),$$

où $C_{pa}^{pa} = \sum_{q=1}^Q C_{paq}^{pa}$. De cette manière, on est sûr que les estimations obtenues antérieurement à l'aide du modèle empirique de Bayes soumis à la méthode itérative du quotient concordent avec le total de domaine initial. Quant aux autres étapes d'estimation du Programme, elles nécessitent l'utilisation de méthodes démographiques. De fait, un des objectifs de la méthode empirique de Bayes est de produire des estimations initiales pour les méthodes démographiques. Voir Michalowski (1993) pour plus de détails.

soumise à la méthode itérative du quotient et l'estimation empirique de Bayes soumise à la MIQ du nombre de personnes oubliées.

4. CORRECTION DES ESTIMATIONS EMPIRIQUES DE BAYES

4.1 Justification et méthode

L'analyse précédente montre clairement qu'il est avan-

tageux d'utiliser la méthode empirique de Bayes. Toutefois, cette méthode ne préserve pas les estimations directes d'enquête qui sont fiables, c.-à-d. celles qui ont rapport aux grands domaines; c'est donc dire que les estimations directes des totaux provinciaux et des totaux par groupe d'âge-sexe ne concordent pas avec les estimations empiriques de Bayes correspondantes. Comme les deux enquêtes mentionnées plus haut ont été conçues pour produire des estimations au niveau des grands domaines, il est indispensable que les estimations empiriques de Bayes concordent avec les totaux marginaux réputés fiables.

Pour faire en sorte qu'il y ait concordance entre les estimations directes des totaux par province et par groupe d'âge-sexe et les estimations empiriques de Bayes finales correspondantes, nous avons eu recours à la méthode itérative du quotient. Il s'agit en fait de la méthode utilisée en Australie pour calculer les estimations pour petits domaines (voir Choi et coll. 1988). Cette méthode consiste à rajuster les estimations empiriques de Bayes de manière qu'elles concordent avec les totaux officiels pour chaque province et pour chaque groupe d'âge-sexe à l'échelle nationale. Une fois la convergence réalisée, les estimations finales sont conformes aux totaux établis directement par l'enquête. Dans la perspective d'un modèle log-linéaire, nous dirons que nous utilisons les résultats des deux études relatives à la couverture comme estimations des effets principaux (province et groupe d'âge-sexe) et les résultats du modèle empirique de Bayes comme estimations des termes d'interaction (province par groupe d'âge-sexe).

On peut décrire la méthode de la façon suivante. Supposons une matrice du nombre estimé de personnes oubliées qui compte P colonnes (pour les provinces) et A lignes (pour les groupes d'âge-sexe). Posons d'abord $F_{pa} = F_i$, puis posons $M_{pa} = C_{pa}(F_{pa} - 1)$, comme l'estimation directe du nombre de personnes oubliées pour la province p et le groupe d'âge-sexe a et $M_{pa}^{(0)} = C_{pa}(F_{pa}^{(0)} - 1)$ comme l'estimation empirique de Bayes du nombre de personnes oubliées, d'après le modèle du même nom. En supposant que le signe (+) signifie l'addition des valeurs de la variable, on peut alors exprimer l'estimation de la méthode itérative du quotient (MIQ) pour les cycles $k = 0, 1, \dots$ par les formules:

$$M_{pa}^{(2k+1)} = M_{pa}^{(2k)} \left(\sum_A M_{pa}^s / \sum_A M_{pa}^{(2k)} \right)$$

et

$$M_{pa}^{(2k+2)} = M_{pa}^{(2k+1)} \left(\sum_P M_{pa}^s / \sum_P M_{pa}^{(2k+1)} \right).$$

$$\widehat{EQM}[F_{pb}'] = \widehat{EQM}(F_{pb}') + 2\delta_{3i}(f_2^2) = \delta_{1i}(f_2^2) + \delta_{2i}(f_2^2) + 2\delta_{3i}(f_2^2).$$

se calcule par la formule

tandis que, d'après la section 2.3, l'EQM de Prasad-Rao

$$\log \hat{v}(F_i) = -6.133 - 0.285 \log C_i,$$

au moyen de la formule

L'objectif du modèle empirique de Bayes est de produire des estimations qui ont une EQM moins élevée que les estimations d'enquête. D'après la section 2.2, on peut montrer que la variance des estimations d'enquête se calcule

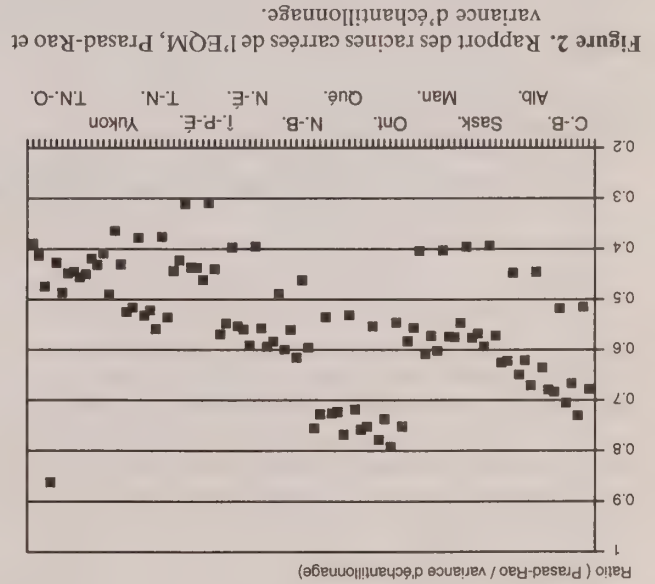


Figure 2. Rapport des racines carrées de l'EQM, Prasad-Rao et variance d'échantillonnage.

Le graphique de la figure 2 reproduit pour chaque domaine la valeur de $R = \sqrt{EQM[F_{pb}'] / \hat{v}(F_i)}$, c'est-à-dire la racine carrée du rapport de l'erreur quadratique moyenne de l'estimation empirique de Bayes à la variance d'échantillonnage estimée (notons que, pour chaque province, les domaines sont classés dans l'ordre suivant: hommes de 0 à 19 ans, de 20 à 29 ans, de 30 à 44 ans, de 45 ans et plus; femmes de 0 à 19 ans, de 20 à 29 ans, de 30 à 44 ans, de 45 ans et plus). De toute évidence, l'EQM de l'estimation empirique de Bayes est plus petite que la variance d'échantillonnage estimée comme l'Ontario et le Québec, le rapport est plutôt élevé, avec des valeurs se situant entre 0.7 et 0.8. Ce gain relativement faible reflète l'importance de la taille des échantillons tirés de ces domaines, lesquels échantillons permettent par ailleurs une estimation fiable de la variance. Les gains les plus appréciables sont enregistrés dans les petites provinces et les territoires. Pour l'Île-du-Prince-Édouard par exemple, le rapport est inférieur à 0.5 dans tous les cas, ce qui montre combien l'amélioration des estimations a été notable. Il y a une valeur aberrante et elle concerne les Territoires du Nord-Ouest (femmes de 0 à 19 ans): l'EQM de Prasad-Rao semble avoir été surestimée pour ce domaine.

Estimations finales des variables utilisées dans la régression

Catégorie	Variable	Estimation finale (valeur absolue)	(β)	($H_0: \beta = 0$)
Moyenne	Moyenne	575,72	1,0075	
Groupe d'âge-sexe	Hommes de 20 à 29 ans	15,34	0,0563	
	Hommes de 30 à 44 ans	5,81	0,0208	
	Hommes de 20 à 29 ans	6,49	0,0240	
Sexe x âge x langue non officielle	Femmes – langue – 0 à 19 ans	2,75	0,0797	
Mode d'occupation	Localités – Colombie-Britannique	3,96	0,0449	
	Localités – Ontario	7,35	0,0804	
	Localités – Québec	2,66	0,0255	
	Localités – Nouveau Brunswick	5,61	0,1064	
	Localités – Yukon	3,80	0,0639	
	Localités – Territoires du Nord-Ouest	6,22	0,0682	

valeurs aberrantes: l'analyse n'a révélé aucun écart notable par rapport à la distribution hypothétique des résidus. Pour plus de détails sur l'analyse des résidus, voir Dick (1993).

Tableau 2

Facteurs de redressement estimés – Estimations directes, estimations lissées et estimations de la méthode itérative du quotient

Sexe	Âge	Estimation	C.-B.	Alb.	Sask.	Man.	Ont.	Qué.	N.-B.	N.-É.	I.P.-É.	T.-N.	Yukon	T.N.-O.
Hommes	0-19	Directe	1,017	1,026	1,012	1,029	1,028	1,017	1,022	1,019	1,004	0,999	1,031	1,036
		Lissée	1,019	1,013	1,016	1,013	1,029	1,016	1,027	1,010	1,007	1,006	1,026	1,027
		MIQ	1,020	1,016	1,011	1,015	1,031	1,018	1,027	1,013	1,005	1,007	1,031	1,036
	20-29	Directe	1,087	1,036	1,068	1,058	1,113	1,071	1,122	1,063	1,060	1,057	1,098	1,127
		Lissée	1,086	1,056	1,062	1,062	1,104	1,074	1,103	1,064	1,063	1,062	1,094	1,122
		MIQ	1,083	1,061	1,073	1,067	1,101	1,079	1,096	1,073	1,041	1,074	1,096	1,127
	30-44	Directe	1,031	1,021	1,028	1,034	1,054	1,047	1,043	1,018	1,025	1,026	1,069	1,080
		Lissée	1,039	1,026	1,028	1,030	1,053	1,041	1,046	1,026	1,028	1,028	1,052	1,059
		MIQ	1,038	1,028	1,032	1,032	1,051	1,043	1,029	1,018	1,009	1,009	1,053	1,059
	45 +	Directe	1,019	1,018	1,002	1,014	1,013	1,011	1,014	1,016	1,018	1,016	0,992	1,076
		Lissée	1,017	1,011	1,006	1,009	1,019	1,016	1,019	1,010	1,009	1,009	1,021	1,039
		MIQ	1,014	1,010	1,006	1,009	1,016	1,012	1,015	1,010	1,005	1,010	1,019	1,035
Femmes	0-19	Directe	1,034	1,018	1,017	1,012	1,037	1,029	1,029	1,014	0,995	1,016	1,026	1,054
		Lissée	1,030	1,015	1,013	1,015	1,038	1,023	1,030	1,010	1,006	1,010	1,028	1,061
		MIQ	1,032	1,018	1,016	1,017	1,040	1,026	1,030	1,012	1,004	1,013	1,030	1,068
	20-29	Directe	1,068	1,047	1,028	1,020	1,043	1,070	1,030	1,004	1,004	1,041	1,068	1,072
		Lissée	1,058	1,036	1,031	1,029	1,044	1,071	1,031	1,027	1,033	1,033	1,069	1,092
		MIQ	1,058	1,041	1,036	1,032	1,048	1,068	1,037	1,018	1,004	1,041	1,072	1,099
	30-44	Directe	1,013	1,009	1,004	1,006	1,027	1,031	1,019	1,004	1,024	1,031	1,020	1,026
		Lissée	1,018	1,008	1,007	1,007	1,030	1,017	1,029	1,010	1,007	1,011	1,028	1,026
		MIQ	1,017	1,008	1,007	1,007	1,028	1,017	1,025	1,011	1,004	1,012	1,027	1,026
	45 +	Directe	1,007	1,003	1,018	1,001	1,011	1,000	1,002	0,993	1,013	1,024	1,007	1,007
		Lissée	1,014	1,006	1,010	1,006	1,021	1,015	1,020	1,006	1,009	1,031	1,031	1,026
		MIQ	1,008	1,004	1,007	1,004	1,012	1,009	1,011	1,004	1,006	1,019	1,016	1,016

3.3 Valeurs estimées des facteurs de redressement

Le tableau 2 donne les estimations directes et les estimations empiriques de Bayes lissées des facteurs de redressement. Un examen attentif du tableau révèle que les deux séries d'estimations sont assez comparables, ce qui dénote l'essence de la méthode empirique de Bayes, qui est de combiner l'estimation directe et l'estimation de modèle. Notons que par suite de l'application de la méthode empirique de Bayes, les valeurs estimées des facteurs de redressement qui, à l'origine, étaient inférieures à un (donc synonymes de surdéveloppement dans certains domaines) sont toutes devenues supérieures à un (donc synonymes de sous-développement). L'écart entre les deux séries d'estimations de facteurs de redressement – en valeur absolue – est inférieur à 1%, et il est même inférieur à 0,5% pour les provinces les plus grandes. Toutefois, pour les territoires et certaines provinces parmi les plus petites, l'écart peut être beaucoup plus grand. Pour les Territoires du Nord-Ouest, l'écart entre l'estimation directe du facteur de redressement et l'estimation empirique de Bayes correspondante est d'environ 2% pour trois groupes d'âge-sexe et il est supérieur à 3% pour le quatrième groupe.

été estimées à l'aide de la méthode exposée dans la section 2. Le modèle linéaire ainsi ajusté comprenait les variables explicatives suivantes:

- a) Une variable indicatrice pour la province ou le territoire.
 - b) Une variable indicatrice pour le sexe.
 - c) Une variable indicatrice pour le groupe d'âge.
 - d) Une variable indiquant le pourcentage de personnes dans un domaine qui sont locataires.
 - e) Une variable indiquant le pourcentage de personnes dans un domaine qui ne parlent aucune des deux langues officielles.
 - f) Diverses variables d'interaction (province × locataire, province × langue non officielle, âge-sexe × locataire).
- En tout, 42 variables ont été utilisées dans le modèle de régression initial.

Le choix de ces variables a été guidé par les résultats de trois séries d'études antérieures sur la question, à savoir la contre-vérification des dossiers (CVD) (Burgess 1988), les études sur la couverture du recensement de 1991 (Germain et Julien 1993) et l'Enquête postcensitaire (EP) des États-Unis, dont font état Hogan (1992) et Datta et coll. (1992). Les motifs de ce choix sont les suivants.

a) L'indicateur de province ou de territoire a été inclus dans le modèle pour indiquer le degré de difficulté de la collecte de données dans chaque province. Avant le recensement de 1991, on supposait que l'opération de collecte serait plus difficile en Colombie-Britannique et en Ontario, et les rapports de recenseurs tendaient à confirmer cette hypothèse.

- b) Les variables âge et sexe ont été incluses dans le modèle à cause des différences reconnues de taux de sous-dénombrement entre les hommes et les femmes. Selon des études antérieures, il y a eu une hausse marquée du sous-dénombrement chez les personnes dans la vingtaine.
- c) Le mode d'occupation ou, plus précisément, le pourcentage de locataires dans chaque domaine a été inclus dans le modèle d'après l'expérience de l'EP des États-Unis et des résultats d'études de CVD antérieures et pour faire suite à une recommandation du Comité consultatif des méthodes statistiques de Statistique Canada.
- d) La variable relative à l'usage d'une langue non officielle a été introduite pour situer géographiquement les groupes d'immigrants et les groupes minoritaires, qui ont eu tendance dans le passé à être sous-dénumérés dans une plus forte proportion que les autres groupes de la société.
- e) Les termes d'interaction ont été inclus afin d'accroître le pouvoir prédictif du modèle.

La moyenne englobe toutes les variables qui ne sont pas incluses dans le modèle. Notons que puisque des variables indicatrices sont utilisées pour la province, le sexe et la combinaison âge-sexe, il faut exclure une variable pour pas avoir de matrice de plan singulière. De fait, la variable manquante, par exemple l'indicateur de province pour Terre-Neuve, est comprise dans la moyenne.

Nous avons de plus appliqué une contrainte opérationnelle au modèle. En effet, le programme IML de SAS qui a été écrit pour estimer les paramètres se limitait à une

matrice de plan de 4,095 éléments numériques; compte tenu de ce qu'il y avait 96 domaines ou observations, il a fallu limiter le modèle à 42 variables.

3.2 Processus de modélisation

Après que nous avons défini le modèle de régression intégral, avec ses 42 variables explicatives, il fallait un procédé pour éliminer les variables qui n'étaient pas statistiquement significatives. Le procédé choisi consistait à éliminer la variable la moins significative au terme de chaque cycle d'estimation. Ainsi, pour le modèle à 42 variables, la variable "femmes locataires de 0 à 19 ans" a été supprimée parce que la valeur *t* correspondante était de 0.05. Nous avons soumis ensuite le modèle de régression, amputé d'une variable, à un nouveau cycle d'estimation. Encore une fois, la variable la moins significative a été éliminée du modèle. Ce procédé équivalait à la régression multiple descendante, décrite dans Draper et Smith (1966, page 167).

La régression multiple descendante a donc servi à éliminer toutes les variables pour lesquelles la valeur *t* était égale ou inférieure à 2 (en valeur absolue). Or, au moment de l'examen du modèle final, nous avons constaté un problème de multicollinéarité entre la variable indicatrice de certaines provinces et les locataires. Cela suppose que certaines variables explicatives sont fortement corrélées entre elles et que, par conséquent, on ne peut estimer avec précision tous les paramètres du modèle. Judge et coll. (1984, page 459) proposent une règle empirique pour déterminer dans quelles circonstances la multicollinéarité et ses conséquences peuvent poser un problème réel, et cette règle est la suivante: lorsque le coefficient de corrélation simple entre les variables est plus grand que le coefficient de détermination, *R*². Le modèle final avait un coefficient de détermination de 0.85, tandis que les coefficients de corrélation simple entre les variables pertinentes étaient tous supérieurs à 0.90 (en valeur absolue, étant donné que les corrélations étaient négatives).

Une façon de résoudre le problème consistait à supprimer les variables pour lesquelles la valeur *t* était la moins élevée, et ces variables se trouvaient être les indicateurs de provinces. Les résultats du modèle final, notamment les coefficients estimés et les valeurs *t* correspondantes, sont présentés dans le tableau 1. La suppression des indicateurs de provinces a eu pour conséquence de rabaisser la valeur de *R*² de 0.85 à 0.844, donc peu de perte de pouvoir prédictif.

Nous avons ensuite soumis le modèle de régression final à divers tests de diagnostic. Comme ce modèle est une régression par les moindres carrés pondérés avec un terme d'erreur aléatoire, Lange et Ryan (1989) proposent d'utiliser la formule suivante pour obtenir des résidus normalisés:

$$z_i = \frac{F'_{(nb)} - X'_i \hat{\beta}}{\sqrt{\sigma^2 + \frac{1}{n-2}}}$$

Les résidus ont été analysés à l'aide de graphiques Q-Q (pour quantile-quantile) et de méthodes de détection des

dans le modèle empirique de Bayes aurait pour conséquence de donner plus de "crédibilité statistique" au modèle de régression qu'à l'estimation directe d'enquête, ce que nous voulons justement éviter. Dick (1993) montre qu'il y a peu de différence entre les estimations de la variance de modèle calculées selon la méthode MVC et celles calculées selon la MM. Comme la méthode MVC a une théorie asymptotique qui est bien comprise, c'est elle qui a été retenue pour l'estimation de la variance de modèle dans le modèle empirique de Bayes.

Harville décrit en détail l'estimation MVC. Essentiellement, il s'agit d'abord d'estimer le paramètre de régression puis d'estimer la variance de modèle à partir des résidus plutôt que des données réelles. Si nous posons X^* comme une matrice de $(n - p)$ contrastes linéaires définis de telle sorte que $E[X^{*'}F] = 0$, Harville montre dans ces conditions que la fonction de vraisemblance (logarithmique) résultante, L_{MVC} , produira des estimations du maximum de vraisemblance avec contrainte si elle est maximisée par rapport à la variance de modèle inconnue. Dans le contexte du modèle empirique de Bayes, on peut décrire la méthode de Harville de la façon suivante. Premièrement, on calcule une estimation initiale (désignée habituellement par zéro) de la variance de modèle, τ_0^2 ; ensuite, on estime le paramètre de régression, β , par les moindres carrés pondérés:

$$\hat{\beta}_{(1)} = (X'V_0^{-1}X)^{-1}X'V_0^{-1}F, \quad (2)$$

où $V_0 = \text{diag}(\tau_0^2) + \sigma_i^2: i = 1, \dots, n$. À l'aide de l'estimation $\hat{\beta}_{(1)}$, on peut calculer une nouvelle estimation MVC de la variance de modèle, $\tau_{(1)}^2$, par l'équation

$$\tau_{k+1}^2 = \tau_k^2 + \left(\frac{\partial L_{MVC}}{\partial \tau^2} \right) [l(\tau^2)]^{-1}, \quad k = 0, 1, \dots, \quad (3)$$

où, si nous posons $F_k = V_k^{-1} - V_k^{-1}X(X'V_k^{-1}X)^{-1}X'V_k^{-1}$, nous avons

$$\frac{\partial L_{MVC}}{\partial \tau^2} = -\frac{1}{2} \text{trace } F_k + \frac{1}{2} V_k^{-1} (F - X\hat{\beta})' V_k^{-1}$$

et

$$l(\tau^2) = -E \left[\frac{\partial^2 L_{MVC}}{\partial (\tau^2)^2} \right] = \frac{2}{1} \text{trace } (F_k' F_k).$$

Notons que s'il y a convergence de τ^2 et de β , $l(\tau^2) - 1$ sera la variance asymptotique de τ^2 .

En soumettant les équations (2) et (3) à un processus d'itération, nous obtenons de nouvelles estimations de τ^2 qui serviront à mettre à jour la valeur estimée de β , qui servira elle-même à mettre à jour la valeur estimée de τ^2 . L'itération se poursuit jusqu'à ce qu'une convergence acceptable soit réalisée: dans ce cas-ci, nous avons utilisé le critère $((\tau_{k+1}^2/\tau_k^2) - 1) < 10^{-6}$.

Une fois que les valeurs de β , les paramètres de régression, et de τ^2 , la variance de modèle, ont été calculées, on peut déterminer les estimations lissées finales. Maritz et Lwin (1989) montrent que l'on peut exprimer l'estimation empirique de Bayes, ou estimation lissée, au moyen de la formule

$$F_{eb}^i = (1 - \omega_i)X_i'\hat{\beta} + \omega_i F_i,$$

où $\omega_i = \tau^2(\tau^2 + \sigma_i^2)^{-1}$. Il s'agit d'une combinaison de l'estimation initiale et de l'estimation par régression pondérée par leurs variances respectives. L'objectif du modèle de lissage est de produire une série d'estimations dont l'EQM est moins élevée que celle des estimations initiales. Par des arguments asymptotiques, Prasad et Rao (1990) proposent d'utiliser l'estimateur défini ci-après pour l'erreur quadratique moyenne:

$$EQM[F_{eb}^i] = EQM[F_{eb}^i] + \left[\left(\frac{\partial \omega_i}{\partial \tau^2} \right)^2 \omega_i E(\tau^2 - \tau^2)^2 \right].$$

Cressie (1992) suppose que l'erreur quadratique moyenne de l'estimation empirique de Bayes, si on utilise l'estimation du maximum de vraisemblance avec contrainte, est:

$$EQM[F_{eb}^i] = EQM(F_{eb}^i) + 2\hat{g}_{3i}(\tau^2) =$$

$$\hat{g}_{1i}(\tau^2) + \hat{g}_{2i}(\tau^2) + 2\hat{g}_{3i}(\tau^2),$$

où

$$\hat{g}_{1i}(\tau^2) = \tau^2(1 - \omega_i)$$

$$\hat{g}_{2i}(\tau^2) = (1 - \omega_i)^2 X_i'(X'V^{-1}X)^{-1}X_i'$$

et

$$\hat{g}_{3i}(\tau^2) = (1 - \omega_i)^2(\tau^2 + \sigma_i^2) - [l(\tau^2)]^{-1}.$$

L'hypothèse de la normalité de ϵ_i et de δ_i est importante pour le calcul. Notons que la valeur de la variance d'échantillonnage σ_i^2 est supposée connue.

Prasad et Rao interprètent chacune des trois composantes de l'EQM de la façon suivante: $\hat{g}_{1i}(\tau^2)$ est l'estimation bayésienne de la variance, $\hat{g}_{2i}(\tau^2)$ est la partie de la variation attribuable à l'estimation des paramètres de régression et $\hat{g}_{3i}(\tau^2)$ est la partie de la variation qui découle de l'estimation de la partie de la variation attribuable à l'estimation de la variance d'échantillonnage: on ignore dans quelle mesure cette dernière composante influencerait sur la variance totale mais l'absence de cette composante indique clairement que l'EQM est sous-estimée.

3. MODÈLE LINÉAIRE EMPIRIQUE DE BAYES

3.1 Variables explicatives

Le modèle empirique de Bayes que nous avons décrit plus haut a été ajusté aux 96 valeurs de facteurs de redressement observées, les variances d'échantillonnage ayant

L'interprétation que l'on peut donner de ce modèle comporte certaines limites. Premièrement, il faut souligner que ce modèle est purement descriptif; on ne peut le considérer comme un modèle causal, ce qui fait que les inférences relatives aux paramètres de régression, β , n'ont pas une importance primordiale même si elles sont intéressantes. Par conséquent, si le modèle de régression final contient un terme concernant, par exemple, les localités de Colombie-Britannique mais ne contient pas de terme sur les localités du Manitoba, cela signifie simplement que les localités de Colombie-Britannique expliquent une forte proportion de la variation des facteurs de redressement pour la Colombie-Britannique, tandis que les localités du Manitoba n'expliquent pas une forte proportion de la variation des facteurs de redressement pour le Manitoba. Comme nous l'avons dit plus haut, la variance d'échantillonnage de l'estimation directe du facteur de redressement est supposée connue dans le modèle empirique de Bayes. Or, l'expérience montre que les variances de ce genre sont en fait quelque peu instables. Afin d'assurer une certaine stabilité dans l'estimation de ces variances, nous proposons la modélisation. Si nous prenons en considération le plan des deux enquêtes par sondage, nous notons que Dick (1993) a montré, suivant des hypothèses modérées, que pour chaque domaine, la variance de l'estimation du nombre de personnes oubliées est proportionnelle au chiffre du recensement élevé à une puissance. Si nous ajoutons les paramètres de normalisation appropriés, cette relation peut être exprimée comme suit:

$$\sigma_i^2 C_i^2 = V(M_i) = K C_i^\gamma,$$

ou sous la forme d'une équation de régression,

$$\text{Log}(V(M_i)) = \alpha + \gamma \text{Log}(C_i) + \eta_i \quad \text{avec}$$

$$\eta_i \sim N(0, \sigma^2).$$

Ce modèle de la variance d'échantillonnage suppose que le produit de l'effet du plan par le taux de sous-dénombrement est le même pour chaque domaine. D'après le rapport de Dick (1993), cette hypothèse semble raisonnable. La figure 1 contient un graphique qui met en relation, pour les 96 domaines, la variance observée de l'estimation du nombre de personnes oubliées, d'après les deux études sur la couverture, et le chiffre du recensement. La droite de régression des moindres carrés a été estimée au moyen de l'équation

$$\text{Log}(\hat{V}(M_i)) = -6.133 + 1.715 \text{Log} C_i.$$

Cette droite est représentée dans la figure 1. L'analyse des résidus (Dick 1993) ne permet pas de croire que les hypothèses du modèle n'auraient pas été respectées. Comme, en outre, le coefficient de détermination, R^2 , a une valeur de 0.943, on a retenu ce modèle pour la production des variances d'échantillonnage. On s'est servi de l'équation suivante pour calculer les variances estimées pour les facteurs de redressement

$$\hat{V}(F_i) = \hat{V}(M_i)/C_i^2.$$

On sait que l'estimation des composantes de la variance selon la méthode EMV comporte un biais par défaut (Harville 1977). La sous-estimation de la variance de modèle

où la variance d'échantillonnage, σ_i^2 , est encore supposée connue.

$$F_i^{(cb)} = E(\Theta_i | \hat{\beta}, \sigma_i^2, \tau^2, F_i),$$

Or, dans la pratique, la variance de modèle, τ^2 , est également inconnue et doit donc être estimée. L'équation de l'espérance conditionnelle de Θ_i deviendra donc

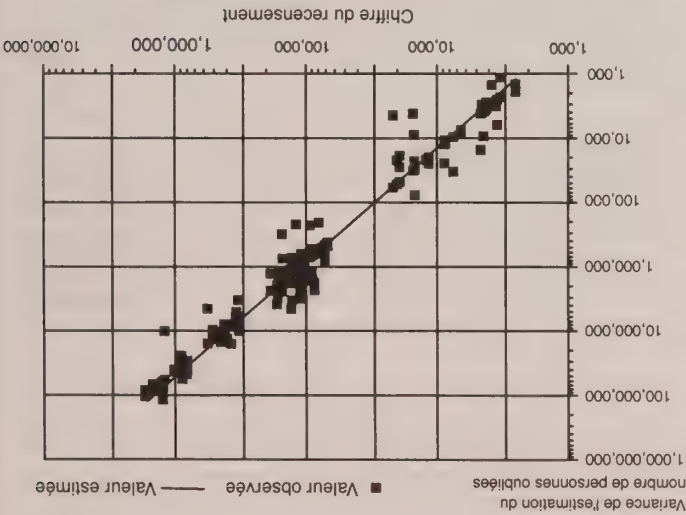
$$\hat{F}_i^{(cb)} = E(\Theta_i | \hat{\beta}, \sigma_i^2, \tau^2, F_i).$$

Jusqu'ici, nous avons décrit le modèle en termes purement bayésiens, c'est-à-dire que seul le paramètre Θ_i est considéré comme inconnu. En adoptant l'approche empirique de Bayes habituelle (Maritz et Lwin 1989), nous supposons connus tous les paramètres, sauf β , le paramètre de régression. L'espérance conditionnelle de Θ_i , étant donné le paramètre de régression estimé, peut être formulée comme suit:

2.3 Estimation de paramètres

On supposera que ces valeurs de la variance d'échantillonnage correspondent aux "variances connues" dont il était question plus haut avec le modèle empirique de Bayes.

Figure 1. Variance observée et chiffre du recensement.



d) les erreurs aléatoires ϵ_i sont distribuées normalement dans chaque domaine.

Nous reviendrons plus loin sur l'hypothèse que la variance d'échantillonnage est connue pour chaque domaine (hypothèse b).

La seconde partie du modèle (modèle de régression) décrit la relation entre les facteurs de redressement vrais et un ensemble de variables explicatives. Ainsi, facteur de redressement vrai = combinaison linéaire de variables explicatives + erreur aléatoire.

Le modèle de régression est formulé par l'équation

$$\Theta_i = X_i \beta + \delta_i : \delta_i \sim \text{Normal}(0, \tau^2),$$

$$i = 1, 2, \dots, n = A \times P,$$

où X_i est la i -ième ligne de X , une matrice ($n \times p$) connue de variables explicatives, β est un vecteur ($p \times 1$) de paramètres de régression inconnus et δ_i est l'erreur aléatoire (différente de ϵ_i) ayant comme variance de modèle τ^2 . Les hypothèses sur lesquelles repose le modèle sont les suivantes:

- les erreurs de modèle, δ_i , ont une moyenne nulle;
- la variance de modèle, τ^2 , est la même pour les n domaines;
- les erreurs de modèle, δ_i , sont distribuées normalement;
- les erreurs de modèle, δ_i , sont indépendantes des erreurs d'échantillonnage, ϵ_i ;
- la covariance des erreurs de modèle de deux domaines différents est nulle (c.-à-d. $\text{Cov}(\delta_i, \delta_j) = 0$).

Il s'agit maintenant d'utiliser le modèle d'échantillonnage et le modèle de régression pour estimer Θ_i , le facteur de redressement vrai. Il est possible de déterminer l'espérance conditionnelle de Θ_i , étant donné $\beta, \sigma_i^2, \tau^2, F_i$ pour le modèle conjoint. À l'aide d'arguments standard (Rao 1973), on peut montrer que l'espérance conditionnelle de Θ_i est:

$$E(\Theta_i | \beta, \sigma_i^2, \tau^2, F_i) = (1 - \omega_i) X_i \beta + \omega_i F_i, \quad (1)$$

$$\text{où } \omega_i = \tau^2(\tau^2 + \sigma_i^2)^{-1}.$$

L'équation (1) est à la base de toutes les estimations qui vont suivre, bien qu'il faille lui apporter quelques modifications avant de l'appliquer aux données. Notons que cette équation est essentiellement une moyenne pondérée de l'estimation directe et de l'estimation par modèle de régression du facteur de redressement. Chaque estimation est pondérée selon le degré de précision avec lequel elle a été calculée. Si l'erreur d'échantillonnage, σ_i^2 , est faible par rapport à l'erreur de modèle, τ^2 , ce qui suppose que l'estimation directe d'enquête est relativement précise, l'estimation lissée finale sera surtout déterminée par l'estimation directe. Par contre, si la variance d'échantillonnage est plus forte que la variance de modèle, l'estimation lissée finale sera déterminée principalement par le meilleur prédicteur linéaire sans biais. La mesure dans laquelle chaque estimation influence sur l'estimation lissée finale est réglée par le coefficient de pondération, ω_i .

Le taux de sous-dénombrement (U_i), que l'on trouve habituellement dans les rapports de Statistique Canada, est lié au facteur de redressement par la relation

$$U_i = M_i(M_i' + C_i)^{-1} = 1 - \Theta_i^{-1}.$$

La modélisation des facteurs de redressement exige la création d'un minimum de domaines essentiels. Ce sont les domaines pour lesquels seront produites des estimations directes des facteurs de redressement. Il doit y avoir une estimation pour chaque province (10) et chaque territoire (2); la valeur de P est donc fixée à 12. Le nombre de groupes d'âge a été fixé à 4 pour obtenir des estimations nationales qui aient une erreur type raisonnablement faible. Ces groupes d'âge sont définis comme suit, pour les hommes et les femmes: de 0 à 19 ans; de 20 à 29 ans; de 30 à 44 ans; et 45 ans et plus. Cela fait donc en tout 96 (12×8) estimations directes de facteurs de redressement à introduire dans le modèle empirique de Bayes. Outre une estimation directe du facteur de redressement, il faut, pour chaque domaine, une estimation de la variance d'échantillonnage correspondante.

2.2 Modèle et hypothèses

Le modèle de base pour le sous-dénombrement se compose de deux parties. La première partie décrit la relation entre les estimations directes d'enquête et les facteurs de redressement vrais, tandis que la seconde décrit la relation entre les facteurs de redressement vrais et un ensemble de variables explicatives. Comme l'estimation des paramètres du modèle de régression se fait d'abord en estimant les paramètres d'une distribution *a priori* hypothétique puis en tenant les valeurs estimées de ces paramètres pour acquises dans les calculs ultérieurs, on parle alors d'un modèle empirique de Bayes (Maritz et Lwin 1989).

La première partie du modèle (modèle d'échantillonnage) décrit le lien entre les facteurs de redressement observés et les facteurs de redressement vrais. Cette relation est supposée vraie pour chaque domaine et elle peut être exprimée comme suit:

$$F_i = \Theta_i + \epsilon_i : \epsilon_i \sim \text{Normal}(0, \sigma_i^2),$$

$$i = 1, 2, \dots, n = A \times P,$$

où Θ_i est le facteur de redressement vrai et ϵ_i , la composante d'erreur aléatoire ayant comme variance σ_i^2 . Les hypothèses de ce modèle sont les suivantes:

- les erreurs d'échantillonnage, ϵ_i , ont une moyenne nulle;
- les variances d'échantillonnage, σ_i^2 , sont connues pour chacun des n domaines;
- comme l'échantillonnage s'est fait de façon indépendante dans chaque domaine, la covariance des erreurs d'échantillonnage ϵ_i , pour le domaine i , et ϵ_j , pour le domaine j , est nulle;

les deux enquêtes par sondage et présentons le modèle empirique de Bayes. Nous étudions aussi les hypothèses et les limites du modèle et traitons brièvement l'estimation des paramètres. Dans la section 3, nous présentons les variables explicatives utilisées dans le modèle de régression et décrivons le processus de modélisation; nous présentons aussi le modèle final et ses résultats. Dans la section 4, nous discutons du bien-fondé de l'opération qui consiste à faire concorder les estimations empiriques de Bayes avec des totaux marginaux fiables, puis nous présentons les estimations redressées finales. Enfin, dans la section 5 nous présentons nos conclusions et proposons des pistes de recherche.

2. MODÈLE RELATIF AUX FACTEURS DE REDRESSEMENT

2.1 Renseignements de base et notation

Le modèle relatif aux facteurs de redressement nécessite des données initiales. Ces données proviennent de deux études sur la couverture du recensement: la contre-vérification des dossiers (CVD) et l'étude du surdénombrement (ES). La CVD sert à estimer le nombre de personnes oubliées au recensement, tandis que l'ES sert à estimer le nombre de personnes comptées par erreur. Ces études visent à produire des estimations fiables du sous-dénombrement net pour toutes les provinces, les régions métropolitaines les plus importantes et certains grands domaines à l'échelle nationale, comme le groupe des hommes de 20 à 24 ans. Comme les deux études sont indépendantes, on peut supposer que la variance de l'estimation du sous-dénombrement net sera égale à la somme des variances estimées de la CVD et de l'ES. Pour plus de détails sur ces études, le lecteur se référera à l'étude de Germain et Julien (1993) et au rapport technique sur le recensement de 1991 qui porte sur la couverture (Statistique Canada 1994).

On peut définir les domaines à l'étude en divisant l'échantillon en p ($= 1, 2, \dots, P$) provinces et territoires et a ($= 1, 2, \dots, A$) groupes d'âge-sexe; il y a donc $A \times P$ domaines pour lesquels il faut calculer des estimations. Soit C_i le nombre de personnes du domaine i (province - âge) qui ont été dénombrées lors du recensement et T_i l'effectif vrai de ce domaine. Le nombre net de personnes de la case i qui ont été oubliées est $M_i = T_i - C_i$. Le facteur de redressement, Θ_i , est le rapport entre l'effectif vrai d'un domaine et le chiffre du recensement pour ce domaine, tandis que le taux de sous-dénombrement, U_i , l'unité qui est le plus souvent reproduite dans les rapports des études sur la couverture, est le rapport du nombre de personnes oubliées à la population vraie.

Le facteur de redressement vrai, Θ_i , qui est la variable que nous voulons estimer, peut être exprimé par la formule:

$$\Theta_i = \frac{T_i}{C_i} = \frac{M_i + C_i}{C_i}.$$

Cet article est structuré de la façon suivante. Dans la section 2, nous donnons des renseignements de base sur les estimations démographiques.

Le lecteur trouvera dans Royce (1992) des détails sur les critères techniques qui s'appliquent au redressement des estimations démographiques reconnus (voir Michalowski 1993). Les estimations sont corrigées à leur tour en fonction de principes démographiques (voir Michalowski 1993). Ces dernières estimations du nombre de personnes oubliées. Ces dernières estimations finales servent de base aux estimations pour petit domaine en se servant de la méthode australienne. Les estimations concorder ces estimations avec les totaux marginaux connus du nombre de personnes oubliées. Enfin, on fait quète. On convertit ensuite ces facteurs lissés en estimations EQM moins élevée que les estimations directes d'enquête. Par une estimation empirique de Bayes, on calcule de nouveaux facteurs de redressement qui auront une EQM moins élevée que les estimations directes de redressement vrais. Par une estimation empirique de Bayes, on les estimations directes d'enquête et les facteurs de redressement vrais et un second, pour mettre en rapport l'américaine, un modèle est défini pour les facteurs de Canada de 1991 combine des éléments des méthodes américaines et australiennes. Comme c'est le cas pour la méthode La méthode de lissage proposée pour le recensement du partie des totaux selon l'Etat, ou le Territoire, et le sexe.

L'EP pour les totaux nationaux selon l'âge et le sexe et une du recensement par âge et par sexe avec les estimations de concorder, par un "balayage par itération", les chiffres obtiennent les estimations pour petit domaine en faisant censitaire (EP) et d'une analyse démographique. On dénombrement est estimé au moyen d'une enquête post-inchangés les chiffres officiels du recensement. Le sous-dans les estimations démographiques mais qui laisse les estimations du sous-dénombrement net du recensement Steel et Skinner (1988) décrivent une méthode qui intègre des Américains pour estimer les totaux de domaines. Chai, Les Australiens utilisent une méthode différente de celle du commerce n'a pas donné suite à cette proposition.

dénombrement dans le recensement, mais le Département démographiques postcensitaires en fonction du sous-ailleurs, on avait proposé aussi de redresser les estimations de the Census, a décidé de ne pas redresser les chiffres du des États-Unis, qui est l'organisme d'attaché du Bureau méthode. En juillet 1991, le Département du commerce phiques. Or Freedman et Navidi (1992) ont critiqué cette les estimations directes pour les grandes régions géographiques. Or Freedman et Navidi (1992) ont critiqué cette les estimations lissées de manière qu'elles concorderaient avec Ericksen et Kadane (1985), puis de "balayer par itération" à méthode empirique de Bayes, comme celui proposé par de lisser les estimations directes par un modèle de régression estimés renfermaient une erreur type élevée, on a proposé somes oubliées. Comme certains de ces 1,392 facteurs rajuster les chiffres du recensement en fonction des perfacteurs de redressement estimés serviraient ensuite à pour 1,392 strates a posteriori définies à l'avance. Les (rapport de la population vraie à la population dénombrée) les chiffres du recensement par des facteurs de redressement (Hogan 1992). Initialement, on avait l'intention de multiplier enquête postcensitaire (the Post Enumeration Survey) pour le recensement de 1990 a été estimé au moyen d'une

Modélisation du sous-dénombrement net dans le recensement du Canada de 1991

PETER DICK¹

RÉSUMÉ

En 1991, Statistique Canada a, pour la première fois, redressé les données du Programme des estimations démographiques pour tenir compte du sous-dénombrement dans le recensement de 1991. Les études sur le champ d'observation d'âge-sexe largement définies à l'échelle nationale. Or, pour redresser les séries d'estimations démographiques, il fallait connaître les estimations du sous-dénombrement pour des groupes d'âge-sexe définis à l'intérieur des provinces ou des territoires. Comme les estimations directes d'enquête pour quelques-uns de ces petits domaines avaient une erreur type élevée à cause du trop petit nombre d'unités de l'échantillon contenues dans ces domaines, il a fallu recourir à des méthodes de modélisation pour petite région. Afin de tenir compte de la plage de variabilité des estimations directes d'enquête, on s'est servi d'un modèle de régression fondé sur une méthode empirique de Bayes pour estimer le sous-dénombrement dans les petits domaines. On a ensuite appliqué la méthode itérative du quotient aux estimations du sous-dénombrement pour qu'il y ait concordance avec les estimations directes marginales. Cet article donne les résultats de la modélisation et la réduction estimée de l'erreur type.

MOTS CLÉS: Petite région; méthode empirique de Bayes; sous-dénombrement.

1. INTRODUCTION ET CONTEXTE

Le recensement du Canada a lieu à tous les cinq ans; un de ses objectifs est de fournir au Programme des estimations démographiques des chiffres de population exacts par groupe d'âge-sexe dans chaque province et territoire. Malheureusement, il y a des personnes qui ne sont pas dénombrées au cours d'un recensement. Dans son évaluation du recensement, Statistique Canada estime le nombre net de personnes oubliées à l'aide de deux enquêtes par sondage, soit la contre-vérification des dossiers, qui sert à estimer le nombre brut de personnes oubliées dans le recensement, et l'étude du surdénombrement, qui sert à estimer le nombre de personnes qui ont été comptées deux fois ou incluses par erreur dans le total. Les estimations des deux enquêtes combinées donnent le nombre net de personnes oubliées dans le recensement. Ces enquêtes sont conçues pour produire des estimations directes fiables pour de grandes régions, comme les provinces, et de grands domaines, comme les groupes d'âge-sexe au niveau national. Or le Programme des estimations démographiques exige des estimations du nombre de personnes oubliées par âge et par sexe pour chaque province. Donc, si on utilisait les estimations directes d'enquête à cette fin, on obtiendrait des estimations qui auraient une erreur type excessivement élevée à cause du trop petit nombre d'unités de l'échantillon contenues dans les petits domaines. Une façon de réduire la variance de ces estimations serait d'emprunter de l'information à des domaines voisins. Cette approche suppose la création d'un modèle explicite pour le petit domaine, qui peut servir à estimer le nombre net de personnes de ce domaine qui ont été oubliées.

La modélisation des estimations pour petit domaine a pour conséquence de produire une série d'estimations qui ont une erreur quadratique moyenne moins élevée que l'estimation directe. Toutefois, contrairement à l'estimation directe d'enquête, qui est non biaisée selon le plan, l'estimation modélisée renfermera un biais. Donc, lorsqu'on veut modéliser les estimations pour petit domaine, il faut être prêt à accepter l'introduction d'un biais dans chaque estimation en échange d'une réduction de la variance. Mais pour faire en sorte que les estimations directes d'enquête, qui sont plus justes, soient utilisées, on recourt à un modèle empirique de Bayes. On obtient ainsi une estimation qui est la combinaison d'une estimation de modèle et d'une estimation directe d'enquête pondérées par leurs variances respectives. On l'appelle estimation empirique de Bayes plutôt que estimation de Bayes parce que les paramètres pertinents sont tout d'abord estimés, puis ces paramètres sont tenus pour acquis dans les calculs ultérieurs. Notons que comme la variance d'échantillonnage entre dans le calcul de l'estimation empirique de Bayes, l'estimation directe aura beaucoup plus de poids dans le résultat final si elle est plus précise. De cette manière, les estimations de modèle ne peuvent supplanter des estimations déjà jugées fiables. La question peut aussi être résolue à l'aide de la méthode hiérarchique de Bayes; pour plus de détails sur cette méthode, voir Datta et coll. (1992). Ghosh et Rao (1994) font une évaluation des deux méthodes, hiérarchique et empirique, dans le contexte de l'estimation pour petit domaine. À l'extérieur du Canada, la littérature statistique décrit deux méthodes de lissage du sous-dénombrement dans le recensement. Aux États-Unis, le sous-dénombrement net

¹ Peter Dick, Division des méthodes d'enquêtes sociales, Statistique Canada, Ottawa (Ontario) Canada, K1A 0T6.

- LEMAÎTRE, G., et DUFOUR, J. (1987). Une méthode intégrée de pondération des personnes et des familles. *Techniques d'enquête*, 13, 211-220.
- LEPKOWSKI, J. (1989). Treatment of wave nonresponse in panel surveys. Dans *Panel Surveys*, (Éds. D. Kasprzyk, G. Duncan, G. Kalton et M.P. Singh). New York: Wiley, 348-374.
- LEPKOWSKI, J.M., MILLER, D.P., KALTON G., et SINGH, R. (1993). Imputation pour la non-réponse de vague dans l'enquête "Survey of Income and Program Participation". *Recueil: Symposium 92, conception et analyse des enquêtes longitudinales*, Statistique Canada, 115-126.
- LITTLE, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *Revue Internationale de Statistique*, 54, 139-157.
- LITTLE, R.J.A. (1989). Sampling weights in the PSID: Issues and comments. Panel Study of Income Dynamics Working Paper, Ann Arbor: University of Michigan.
- NELSON, D., MCMILLEN, D., et KASPRZYK, D. (1985). *An Overview of the SIPP, Update I*. SIPP Working Paper No. 8401. Washington D.C.: U.S. Bureau of the Census.
- SINGH, R., HUGGINS, V., et KASPRZYK, D. (1990). *Handling Single Wave Nonresponse in Panel Surveys*. SIPP Working Paper No. 9009. Washington D.C.: U.S. Bureau of the Census.
- ZIESCHANG, K.D. (1990). Sample weighting methods and estimation of totals in the Consumer Expenditure Survey. *Journal of the American Statistical Association*, 85, 986-1001.

DEVILLE, J.-C., SÄRNDAAL, C.-E., et SAUTORY, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.

DUNCAN, G.J., et HILL, M.S. (1985). Conceptions of longitudinal households: Fertile or futile? *Journal of Economic and Social Measurement*, 13, 361-375.

ERNST, L.R. (1989). Weighting issues for longitudinal household and family estimates. Dans *Panel Surveys*, (Eds. D. Kasprzyk, G. Duncan, G. Kalton et M.P. Singh). New York: John Wiley, 139-159.

GAILLY, B., et LAVALLÉE, P. (1993). *Insérer des Nouveaux Membres dans un Panel Longitudinal de Ménages et D'Individus: Simulations*. Walferdange, Luxembourg: CEPS/Instead.

HILL, M.S. (1992). *The Panel Study of Income Dynamics: A User's Guide*. Newbury Park, CA: Sage Publications.

HILL, M.S. (1995). Communication personnelle.

HOLT, D., et SMITH, T.M.F. (1979). Post stratification. *Journal of the Royal Statistical Society, A*, 142, 33-46.

HUANG, H. (1984). Obtaining cross-sectional estimates from a longitudinal survey: Experiences of the Income Survey Development Program. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 670-675.

JABINE, T.B., KING, K.E., et PETRONI, R.J. (1990). *Survey of Income and Program Participation: Quality Profile*. Washington D.C.: U.S. Bureau of the Census.

JUDKINS, D., HUBBLE, D., DORSCH, J., MCMILLEN, D., et ERNST, L. (1984). Weighting of persons for SIPP longitudinal tabulations. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 676-687.

KALTON, G. (1986). Handling wave nonresponse in panel surveys. *Journal of Official Statistics*, 2, 303-314.

KALTON, G., et KASPRZYK, D. (1986). Le traitement des données d'enquête manquantes. *Techniques d'enquête*, 12, 1-17.

KASPRZYK, D. (1988). *The Survey of Income and Program Participation: An Overview and Discussion of Research Issues*. SIPP Working Paper No. 8830. Washington D.C.: U.S. Bureau of the Census.

KASPRZYK, D., et MCMILLEN, D.B. (1987). SIP: Characteristics of the 1984 Panel. *Proceedings of the Social Statistics Section, American Statistical Association*, 181-186.

KISH, L. (1965). *Survey Sampling*. New York: Wiley.

LAVALLÉE, P. (1995). Pondération transversale des enquêtes longitudinales menées auprès des individus ou des ménages à l'aide de la méthode du partage des poids. *Techniques d'enquête*, 21, 27-35.

LAVALLÉE, P., et HUNTER, L. (1992). Méthodes de pondération pour l'enquête sur la dynamique du travail et du revenu. *Recueil: Symposium 92, conception et analyse des enquêtes longitudinales*, Statistique Canada, 77-88.

LAVALLÉE, P., MICHAUD, S., et WEBBER, M. (1993). The Survey of Labour and Income Dynamics, design issues for a new longitudinal survey in Canada. *Bulletin de l'Institut International de Statistique*, 49ème Session, Communication libres, livre 2, 99-100.

REMERCIEMENTS

La classe des méthodes de pondération décrite ici a une gamme d'applications plus vaste que celles qui ont été présentées. Elle peut en fait se révéler utile dans tous les cas où une pondération selon les probabilités de sélection inverses aurait été idéale, mais où les probabilités d'inclusion et les probabilités d'inclusion conjointe ne sont pas toutes connues. Prenons, par exemple, la version modifiée, décrite par Brick et Wakseberg (1991), de la méthode d'enquête téléphonique au hasard de Mitofsky-Wakseberg. Un échantillon de numéro de téléphone (numéros de base) est sélectionné au premier degré de ce plan d'échantillonnage à deux degrés. Si un des numéros de base est un numéro résidentiel en service, le ménage est sélectionné et un nombre fixe de numéros de téléphone additionnels est sélectionné dans le même groupe de 100 numéros. Les ménages trouvés à ces numéros sont alors tous inclus dans l'échantillon. Si un numéro de base n'est pas un numéro en service, le processus d'échantillonnage s'arrête. Selon cette méthode, la probabilité qu'un numéro résidentiel en service soit sélectionné dépend du nombre de numéros résidentiels en service dans son groupe de 100 numéros, et diffère donc d'un groupe à l'autre. Cette probabilité peut être estimée d'après l'échantillon de numéros de téléphone appartenant au groupe de 100. Une complication survient, toutefois, quand un ménage de l'échantillon a deux numéros de téléphone ou plus. Dans ce cas, on peut estimer la probabilité de sélection du numéro de téléphone de l'échantillon, mais pas celles des autres numéros. Par conséquent, la méthode de pondération habituelle selon les probabilités de sélection inverses ne peut s'appliquer. Cependant, la méthode de pondération de remplacement décrite ici peut être utilisée.

BIBLIOGRAPHIE

ALEXANDER, C.H. (1987). Une classe de méthodes utilisant des chiffres de population dans la pondération des ménages. *Techniques d'enquête*, 13, 193-209.

BRICK, J.M., et WAKSEBERG, J. (1991). Méthode pour éviter l'échantillonnage progressif dans une enquête téléphonique à composition aléatoire. *Techniques d'enquête*, 17, 31-46.

BUCK, N., GERSHUNY, J., ROSE, D., et SCOTT, J. (Eds.) (1994). *Changing Households: The British Household Panel Survey 1990-1992*. Colchester, U.K.: ESRC Research Centre on Micro-social Change.

CITRO, C.F., et KALTON, G. (1993). *The Future of the Survey of Income and Program Participation*. Washington D.C.: National Academy Press.

principale, et le poids entièrement rajusté de cette personne principale, est considéré comme le poids du ménage. Puisque les poids des personnes sont déjà rajustés en fonction des totaux de contrôle, le poids du ménage intègre un certain rajustement contribuant à réduire le biais dû au sous-dénombrement. Dans le cas d'estimations transversales à partir d'une enquête à panel de ménages, on peut appliquer directement la méthode de la personne principale, conjointement avec les méthodes de pondération égale des ménages et des personnes, pour produire les poids au niveau des ménages. La méthode de la personne principale présente un inconvénient, du fait que les effets calculés d'après les poids des personnes principales diffèrent en général des totaux de contrôle. Les estimations peuvent aussi varier sensiblement selon les critères de détermination de la personne principale du ménage. Cette méthode a également été critiquée parce que les poids des membres du même ménage diffèrent, bien qu'ils aient tous été sélectionnés selon la même probabilité que le ménage. Les méthodes d'estimation proposées par Alexander (1987), Lemaître et Dufour (1987), et Zieschang (1990) tentent de remédier à ces objections en imposant une cohérence entre les poids des ménages et les totaux indépendants au niveau des personnes, et en minimisant la distance entre les poids des ménages initiaux et les poids rajustés. Les trois articles examinent des variantes d'un algorithme des moindres carrés généralisés (MCG) pour atteindre cet objectif. Zieschang (1990) montre comment les MCG peuvent servir à créer des poids qui correspondent aux totaux de contrôle des personnes et qui sont obligatoirement les mêmes pour toutes les personnes d'un même ménage.

Nous avons décrit, dans le présent article, des méthodes de pondération permettant d'effectuer, à des vagues ultérieures d'une enquête à panel de ménages, des analyses transversales utilisant les données sur tous les ménages et toutes les personnes pour lesquels des données sont recueillies. Ces méthodes de pondération peuvent tenir compte des nouveaux venus dans la population qui se joignent aux ménages de membres de la population initiale, mais non des autres nouveaux venus.

7. SOMMAIRE ET CONCLUSION

L'emploi de la méthode de pondération habituelle selon les probabilités de sélection inverses exige de connaître les probabilités de sélection des ménages de tous les membres des ménages inclus dans l'échantillon à une vague ultérieure, ainsi que les probabilités de sélection conjointe des ménages initiaux qui fournissent des membres aux ménages des vagues ultérieures. Il est fréquent que la méthode de pondération selon les probabilités de sélection inverses ne puisse être appliquée, parce que ces probabilités sont inconnues. Pour tenter de remédier à ce problème, nous avons décrit une autre approche, qui exige seulement de connaître les probabilités de sélection des ménages de l'échantillon initial.

Cette autre approche produit une classe de méthodes de pondération, notamment la méthode de pondération égale des personnes (parts équitables) utilisée dans l'enquête SIPP et la méthode de pondération égale des ménages. Toutes les méthodes de cette classe produisent des poids qui, en termes d'espérance, sont égaux à ceux produits par la méthode habituelle fondée sur les probabilités de sélection inverses. La variance des poids autour de ceux produits par les probabilités de sélection inverse entraîne un accroissement de la variance des estimations de l'enquête. Quand les ménages initiaux sont sélectionnés avec des probabilités approximativement égales, la méthode de pondération égale des ménages est quasi optimale, aussi bien pour les analyses au niveau des ménages que pour celles au niveau des personnes, pour contrôler cette augmentation de variance.

Cette autre classe de méthodes produit des estimations sans biais de totaux de la population pour n'importe quel ensemble de constantes α_{ij} qui respecte la condition $\sum_j \alpha_{ij} = 1$ et pour n'importe quel plan de sondage initial. Les méthodes de pondération égale des ménages et des personnes sont, toutefois, sous-optimales pour les échantillons initiaux non sélectionnés avec probabilités égales. L'une d'entre elles peut néanmoins être la méthode qui convient pour de tels plans, car le choix optimal des α_{ij} dépend des probabilités de sélection initiales inconnues, et ne peut donc être déterminé. Les méthodes de pondération égale des ménages et des personnes ont des besoins différents en données, la première exigeant la connaissance du nombre de ménages de la vague t , tandis que la deuxième ne l'exige pas. Le fait que cette information ne soit pas toujours immédiatement accessible plaide en faveur de la méthode de pondération égale des personnes.

Les poids transversaux individuels à une vague parti-culière peuvent servir de poids de départ à une analyse longitudinale qui commence à cette vague. Cette méthode englobe dans l'analyse longitudinale les cohabitants présents à ce moment. Toutefois, si les cohabitants ne sont pas suivis lorsqu'ils cessent de vivre avec les membres de l'échantillon, ceux qui quittent les membres de l'échantillon avant la fin de la période de l'analyse longitudinale deviennent des non-répondants. Avant d'inclure les cohabitants dans une analyse longitudinale, il importe de s'assurer qu'on ne risque pas, ce faisant, d'engendrer un important biais de non-réponse.

Les deux formes de rajustements dus à la non-réponse ci-dessus sont définies par rapport aux ménages initiaux. Un autre type de non-réponse des ménages ne peut être traité de cette façon. Ce type de non-réponse survient lorsqu'un ménage initial se subdivise en deux ménages distincts ou plus à la vague t , et que certains de ces ménages, mais pas tous, répondent à cette vague. Dans ce cas, le rajustement dû aux ménages non-répondants doit être fait par rapport aux ménages de la vague t , c'est-à-dire H_t , plutôt que par rapport aux ménages initiaux H_j . Si le nombre de ménages initiaux ayant des membres dans chaque ménage non-répondant de ce type de la vague t était connu, les poids w_j de ces ménages pourraient être calculés selon l'approche décrite ci-dessus. On pourrait alors appliquer simplement des rajustements de pondération à l'intérieur des classes de pondération des ménages de la vague t , et ainsi compenser la non-réponse des ménages. Toutefois, il arrive souvent qu'en pratique, le nombre de ménages initiaux ayant des membres dans un ménage non-répondant à la vague t ne soit pas connu. Une solution possible, dans ce cas, consiste à estimer ce nombre en le supposant égal au nombre moyen pour les ménages qui ont répondu à la vague t et qui ont les mêmes caractéristiques (p. ex., ils sont aussi le résultat de subdivisions de ménages initiaux), et appartenant à la même classe de pondération, que le ménage non-répondant. En utilisant au besoin de tels nombres estimés, on peut déterminer les poids w_j pour tous les ménages non-répondants de ce type. Des rajustements de pondération courants peuvent alors être appliqués aux ménages qui répondent à la vague t , pour compenser les cas de non-réponse.

Le sous-dénombrement de la population cible est un autre problème non lié à l'échantillonnage auquel on s'est traditionnellement attaqué, dans les enquêtes, en rajustant les poids d'échantillonnage. Par exemple, la stratification a posteriori (voir, par exemple, Holt et Smith 1979) et des méthodes itératives du quotient généralisées (Deville et coll. 1993) sont souvent utilisées pour rajuster les poids, de façon que leur somme corresponde à des totaux tirés de sources indépendantes non soumises au sous-dénombrement. Il se peut que de tels rajustements réduisent aussi les erreurs d'échantillonnage des estimations, quoique la réduction du biais soit souvent plus critique.

Les totaux de contrôle utilisés dans la plupart des enquêtes auprès des ménages sont des effectifs de classes définies par des caractéristiques comme l'âge, le sexe et la race. Cette façon de réduire le biais de sous-dénombrement peut parfaitement suffire quand des estimations au niveau des personnes sont les seuls types de statistiques que l'enquête doit produire. Toutefois, d'autres étapes sont nécessaires pour calculer les poids, et produire des statistiques, au niveau des ménages.

La méthode de la personne principale, décrite par Alexander (1987), est un moyen dont on dispose pour déterminer les poids au niveau des ménages quand les totaux de contrôle se fondent sur des effectifs de personnes. Dans cette méthode, des rajustements de stratification a posteriori sont appliqués au niveau des personnes. Un membre du ménage est alors désigné comme la personne

de type 2). Avec cette dernière solution, on dispose de données sur tous les membres des ménages répondants, ce qui rend inutiles les rajustements au niveau des personnes à l'intérieur des ménages répondants.

Examinons maintenant les problèmes que posent les données manquantes dans les analyses transversales d'une enquête à panel de ménages. Un fichier transversal distinct contenant les données de tous les ménages répondants et de leurs membres (dans lequel on a soit supprimé les ménages, soit imputé des valeurs, pour les réponses manquantes dans le cas de la non-réponse de ménage partielle) peut être créé pour chaque vague. Il faut ensuite apporter des rajustements pour compenser la non-réponse et le sous-dénombrement des ménages et des personnes dans chaque fichier.

Les ménages non-répondants à la vague t peuvent être répartis entre les cas de non-réponse totale et les cas de non-réponse de vague. La non-réponse totale survient, dans une enquête par panel, lorsqu'une unité de l'échantillon omet de fournir des données à l'une ou l'autre vague. Puisqu'il est courant de ne pas suivre les ménages de l'échantillon qui sont non-répondants à la vague initiale, ces ménages et leurs membres sont généralement considérés comme des cas de non-réponse totale. La compensation est alors relativement simple. Les poids de la vague 1 des ménages répondants à la vague initiale peuvent être rajustés par des méthodes courantes de compensation de la non-réponse, et les poids rajustés peuvent être substitués aux probabilités de sélection pour l'établissement des poids transversaux des vagues ultérieures. La plupart des méthodes de compensation de la non-réponse, par exemple, les rajustements par classe de pondération (Kalton et Kasprzyk 1986) et les rajustements fondés sur les tendances de réponse (Little 1986), se fondent sur l'hypothèse que la non-réponse est aléatoire à l'intérieur de classes de pondération ou que les probabilités de réponse à l'intérieur d'une classe peuvent être estimées avec précision. Dans ces conditions, le mécanisme de réponse peut être traité comme un degré additionnel d'échantillonnage. Ainsi, les probabilités de sélection, p_j , qui servent à définir les poids dans l'équation (3.5) peuvent être redéfinies comme le produit des probabilités de sélection et du rajustement au titre de la non-réponse. Par exemple, si des rajustements par classe de pondération sont appliqués, la probabilité de sélection du ménage initial h_j multipliée par le taux de réponse pondéré pour la classe de pondération à laquelle appartient h_j est utilisée au lieu de la valeur originale p_j . Les résultats qui précèdent sont alors valables, selon les poids rajustés qui précèdent compte de la non-réponse totale.

Cette approche peut être étendue aux rajustements de pondération dans le cas des ménages qui répondent à la vague initiale et qui deviennent des ménages non-répondants à la vague t . Les répondants de la vague initiale peuvent alors être répartis en classes de pondération selon les réponses données à cette vague, et les poids des ménages qui mènent à un ou plusieurs ménages répondants à la vague t peuvent être rajustés de nouveau en compensation de ceux qui mènent à des ménages non-répondants à la vague t . Les w_j révisés peuvent alors être employés dans l'équation (3.5) et ce qui en découle.

$$w_i = \sum_k \sum_j w_{ijk}^* / M_i.$$

On peut aussi, pour les α_{ijk} , faire le même choix que dans la méthode de pondération égale des ménages. Soit C_j le nombre des ménages initiaux qui ont des membres dans le ménage H_i au temps t . Alors la somme $\sum_j \sum_k \alpha_{ijk} = 1$ peut être répartie également entre les ménages, chaque membre du ménage initial h_j recevant la valeur $\alpha_{ijk} = 1/C_j M_{ij}$. Alors, pour le ménage initial h_j

$$\sum_{M_{ij}}^k \alpha_{ijk} = 1/C_j.$$

La détermination des α_{ijk} minimisant la variance du total estimé \bar{Y} pour la population des personnes découle directement du développement correspondant fait pour la population des ménages à la section 3. Le total estimé pour la population des personnes est

$$\bar{Y} = \sum_s^i \sum_{N_i}^k w_{ik} X_{ik} = \sum_s^i \sum_{N_i}^k w_i X_{ik},$$

puisque les poids pour chaque personne du ménage de l'échantillon H_i sont les mêmes. Ce total estimé peut être exprimé ainsi:

$$\bar{Y} = \sum_s^i w_i X_i,$$

où $Y_i = \sum_k X_{ik}$ est le total du ménage pour H_i . Ainsi, \bar{Y} peut être exprimé comme un total au niveau des ménages, et les résultats de la section 3 peuvent être appliqués directement.

Prenons l'exemple de la section 3, dans lequel H_i est formé de membres venant de seulement deux ménages initiaux et comprend peut-être aussi un ou plusieurs nouveaux venus. Dans ce cas, le poids des personnes $w_i = \sum_j \sum_k \alpha_{ijk} w_{ijk}$ se réduit à

$$w_i = \left(\sum_k^k \alpha_{ik} \right) w_i^1 + \left(\sum_k^k \alpha_{ik} \right) w_i^2$$

$$= \alpha_i w_i^1 + (1 - \alpha_i) w_i^2,$$

où $\alpha_i = \sum_k \alpha_{ik}$. Comme l'indique l'équation (3.8), la valeur optimale de α_i est

$$\alpha_{oi} = \left(1 + \frac{p_1 - p_2}{p_2 - p_{12}} \right)^{-1}.$$

Les valeurs individuelles α_{ijk} ne sont pas nécessaires pour calculer les w_i ; seuls les totaux des ménages initiaux $\sum_k \alpha_{ijk}$ sont requis. Si des valeurs individuelles sont nécessaires pour les α_{ijk} , on peut tout simplement les attribuer sous forme du quotient $\sum_k \alpha_{ijk} / M_{ij}$.

Comme dans le cas des ménages, la pondération optimale α_{oi} exige la connaissance de p_1 , p_2 et p_{12} . Si ces probabilités sont connues, le poids selon la méthode habituelle des probabilités de sélection inverses w_i^* peut être calculé, et son utilisation sera préférable. Dans le cas d'un échantillon avec probabilités de sélection approximativement égales, la méthode de pondération égale des ménages devrait donner de bons résultats. Toutefois, la méthode de pondération égale des ménages exige de l'information sur le nombre de ménages initiaux fournissant des membres au ménage courant H_i , et, parfois, cette information n'est pas disponible. Comme il a été mentionné à la section 3, on préférera peut-être pour cette raison la méthode de pondération égale des personnes.

5. ANALYSES LONGITUDINALES AU NIVEAU DES PERSONNES

Les enquêtes par panel offrent un avantage clé, celui de permettre des analyses longitudinales fondées sur le lien entre les variables des mêmes unités de l'échantillon mesurées à des moments différents. Puisque toutes les personnes des ménages de l'échantillon initial sont suivies pendant la durée complète du panel ou jusqu'à ce qu'elles sortent de la population de l'enquête, les données qu'elles fournissent peuvent être soumises directement à des analyses longitudinales portant sur n'importe quelle période se situant à l'intérieur de la durée du panel (bien que des ajustements au titre de la non-réponse puissent être nécessaires pour tenir compte de l'érosion du panel). Par exemple, si le panel dure dix ans, les données sur les membres de l'échantillon initial peuvent être analysées de l'année 1 à l'année 10, de l'année 5 à l'année 9, ou pour toute autre période. Les nouveaux venus (p. ex. naissances) peuvent être inclus dans l'analyse pour des périodes commençant après le début du panel, pourvu qu'ils soient traités comme des membres qui sont suivis tout au long du panel, même lorsqu'ils quittent les ménages des membres de l'échantillon initial.

Avec les méthodes de pondération décrites à la section précédente, les cohabitants peuvent être inclus dans les analyses transversales des vagues ultérieures. Ces méthodes de pondération permettent d'obtenir une représentation transversale de la population à n'importe quelle vague du panel (sauf pour les nouveaux venus ne vivant pas avec des membres de la population initiale). Il est alors possible de considérer l'échantillon total formé des membres de l'échantillon initial et des cohabitants au temps t comme l'échantillon initial d'un nouveau panel pouvant servir aux analyses longitudinales allant du temps t au temps $(t + k)$. Cette méthode est utilisée, par exemple, dans l'enquête SIPP: tous les membres de l'échantillon initial et les cohabitants présents au début de la deuxième année du panel sont inclus dans les analyses relatives à cette année-là.

L'inclusion des cohabitants dans l'analyse longitudinale comporte une contrainte du fait que les règles de suivi utilisées dans la plupart des enquêtes par panel précèdent

nouveaux venus qui se joignent à des ménages comprenant un ou plusieurs membres de la population initiale peuvent être représentés dans les estimations transversales des vagues ultérieures, mais les nouveaux venus vivant dans des ménages qui ne comprennent aucun membre de la population initiale ne sont pas couverts (à moins qu'un échantillon spécial puisse être prélevé). Le premier type de nouveaux venus est pris en compte dans la méthode de pondération décrite ci-dessous, mais le deuxième type ne l'est pas.

Supposons qu'il y ait N personnes dans la population au temps t , avec N_t personnes dans le ménage H_i ($i = 1, 2, \dots, H$) et $\sum N_t = N$. Les membres du ménage H_i viennent des ménages h_j, h_k, h_l, \dots , qui existaient au temps t . Soit M_{ij} le nombre de membres du ménage H_i au temps t qui appartenaient au ménage h_j au début du panel. La somme $M = \sum M_{ij}$ est inférieure à la taille de la population au temps t parce que des personnes sont sorties de la population entre le temps t et le temps t , et $M < N$ en raison de la présence, dans la population, de nouveaux venus qui vivent dans des ménages comprenant des membres de la population initiale.

Examinons maintenant l'estimation d'un total pour la population de personnes au temps t :

$$Y = \sum_H \sum_{N_t}^{k=1} Y_{tk}. \quad (4.1)$$

$$Y = \sum_H \sum_{N_t}^{k=1} w_{tk} Y_{tk}, \quad (4.2)$$

où w_{tk} est la valeur pour la personne k dans le ménage H_i . Comme dans l'analyse au niveau des ménages décrite dans la section précédente, un estimateur général pour ce total peut s'écrire

que $E(w_{tk}) = 1$ pour tous les t et tous les k . Comme il a été signalé plus haut, il y a de nombreuses façons de remplir la condition $E(w_{tk}) = 1$. Il est instructif d'en examiner trois. Premièrement, posons $w_{tk} = 0$ pour toutes les personnes ne faisant pas partie de l'échantillon initial. Dans ce cas, l'estimateur Y laisse de côté les cohabitants. Soit p_{tk} la probabilité qu'un membre de la population initiale, la personne k vivant dans le ménage H_i au temps t , soit sélectionnée dans l'échantillon initial, et soit $w_{tk} = 1/p_{tk}$. Alors, pour une telle personne

$$E(w_{tk}) = p_{tk}(1/p_{tk}) + (1 - p_{tk})0 = 1.$$

Selon cette méthode, pour tous les nouveaux venus dans la population, $w_{tk} = 0$ avec certitude. Ainsi, Y dans (4.2) fournit un estimateur sans biais du total pour la population initiale qui est encore présente au temps t , mais n'inclut pas de composante pour les nouveaux venus.

La méthode ci-dessus peut être modifiée pour tenir compte de certains types de nouveaux venus. Par exemple, on peut inclure les nouveaux-nés des mères faisant partie de l'échantillon en leur attribuant les poids de leurs mères, ou si, comme dans l'enquête SIPP, la population est définie comme étant formée des adultes de 16 ans et plus, on peut traiter les personnes qui ont moins de 16 ans au début du panel comme des membres de l'échantillon et leur attribuer une probabilité, et les inclure dans les analyses lors de vagues ultérieures, après qu'elles ont eu 16 ans. De telles modifications, toutefois, ne tiennent pas compte de tous les types de nouveaux venus. Pourvu que la proportion des autres types de nouveaux venus soit faible, cette lacune pourrait être considérée comme minimale.

La méthode de pondération qui restreint l'analyse aux membres de l'échantillon initial et à certains nouveaux venus bien définis est employée dans l'enquête PSID. Sa déficience réside dans le fait qu'elle n'utilise pas directement les données recueillies pour les cohabitants. Ces données peuvent servir à expliquer la situation vécue par les membres de l'échantillon, mais les cohabitants sont exclus de l'échantillon pour les fins de l'analyse.

Pour pouvoir inclure les cohabitants dans les analyses transversales au temps t , il faut leur attribuer des poids post-tifs. Puisque la probabilité qu'une personne soit prélevée dans l'échantillon est la même que celle du ménage dont elle est membre, des méthodes de pondération pour les analyses transversales au niveau des personnes à la vague t peuvent être tirées directement de celles obtenues à la section 3 pour les ménages. Ici, nous allons exposer la stratégie générale d'obtention, pour l'analyse transversale au temps t , de poids fondés seulement sur les probabilités de sélection des membres de l'échantillon initial, ce qui nous évitera le problème soulevé par la méthode des probabilités de sélection inverses, qui a été mentionné à la section 3.

Supposons que I_{ijk} désigne la personne k du ménage initial h_j qui fait maintenant partie du ménage H_i . Soit w_i le poids de chaque membre du ménage H_i pour les analyses transversales au temps t , et soit

$$w_i = \sum_j \sum_k \alpha_{ijk} w_{ijk}^*$$

où $w_{ijk}^* = 1/p_j$ si le ménage h_j était dans l'échantillon initial et $w_{ijk}^* = 0$ sinon. Alors, comme auparavant, $E(w_{ijk}^*) = 1$ pour les membres de la population initiale. On peut tenir compte des nouveaux venus, pour qui $p_j = 0$, en posant $\alpha_{ijk} = 0$. Alors,

$$E(w_i) = \sum_{N_t} \sum_k \alpha_{ijk} E(w_{ijk}^*) = \sum_{M_t} \sum_k \alpha_{ijk} = 1$$

pourvu que $\sum_j \sum_k \alpha_{ijk} = 1$. Si cette condition est remplie, Y est sans biais pour Y .

Un choix naturel pour α_{ijk} est de poser $\alpha_{ijk} = 1/M_i$ pour tous les membres de la population initiale. C'est la méthode de pondération égale des personnes, dans laquelle chaque membre du ménage H_i au temps t (y compris les nouveaux venus) reçoit le poids

En supposant que $\text{Cov}(w_i, w_i' | H_i, H_i') = 0$, on obtient

$$V(X | s) = \sum Y_i^2 V(w_i | H_i) = \sum Y_i^2 [E(w_i^2 | H_i) - w_i^{*2}],$$

vu que, comme il a été indiqué plus haut, $E(w_i | H_i) = w_i^*$. Ainsi, en supposant que $\text{Cov}(w_i, w_i' | H_i, H_i') = 0$, $V(X | s)$, est minimisé quand $E(w_i^2 | H_i)$ est minimisé. Reprenons le cas simple examiné plus haut, dans lequel H_i est composé de membres venant de deux ménages initiaux, et posons $w_i = \alpha_i w_i^1 + (1 - \alpha_i) w_i^2$. On a alors:

$$E(w_i^2 | H_i) =$$

$$\frac{\alpha_i^2 (p_1 - p_{12}) \frac{p_1^2}{2} + (p_2 - p_{12}) \frac{(1 - \alpha_i)^2}{2} p_2^2 + p_{12} \left(\frac{\alpha_i}{1 - \alpha_i} + \frac{p_1}{p_2} \right)}{p_1 + p_2 - p_{12}}.$$

Minimiser $E(w_i^2 | H_i)$ équivaut à minimiser

$$\Delta = (p_1 - p_{12}) p_2^2 \alpha_i^2 + (p_2 - p_{12}) p_1^2 (1 - \alpha_i)^2 + p_{12} [(p_2 - p_1) \alpha_i + p_1]^2.$$

Alors,

$$\frac{\partial \Delta}{\partial \alpha_i} = 2(p_1 - p_{12}) p_2^2 \alpha_i - 2(p_2 - p_{12}) p_1^2 (1 - \alpha_i) + 2p_{12}(p_2 - p_1) [(p_2 - p_1) \alpha_i + p_1].$$

Si l'on résout $\partial \Delta / \partial \alpha_i = 0$ en fonction de α_i on obtient le α_i optimal, c'est-à-dire

$$\alpha_i = \left(1 + \frac{p_2}{p_2 - p_{12}} \right)^{-1} \left(1 + \frac{p_1}{p_2 - p_{12}} \right)^{-1}. \quad (3.8)$$

Si les ménages initiaux sont sélectionnés de façon indépendante, c.-à-d. si $p_{12} = p_1 p_2$,

$$\alpha_i = \left[1 + \frac{p_2(1 - p_1)}{p_1(1 - p_2)} \right]^{-1} = \left[1 + \frac{\psi_1}{\psi_2} \right]^{-1}, \quad (3.9)$$

où $\psi_j = p_j / (1 - p_j)$ est la probabilité que le ménage initial h_j soit sélectionné.

Peu importe que les ménages aient été sélectionnés de façon indépendante ou non, dans le cas spécial d'un échantillon prélevé initialement avec probabilités égales, avec $p_1 = p_2$,

$$\alpha_i = \frac{1}{2}.$$

Par conséquent, pour le cas des deux ménages, si l'échantillon initial est prélevé avec probabilités égales, la méthode

de pondération égale des ménages minimise la variance des poids des ménages autour du poids obtenu selon la méthode des probabilités de sélection inverses.

Le choix optimal de α_i donné par (3.8) exige la connaissance de p_1 , p_2 et p_{12} , et celui donné par (3.9) exige qu'on

sait ces probabilités soient indépendantes. Si ces probabilités étaient connues, le poids habituel selon les probabilités inverses de sélection pourrait être utilisé et serait préférable. Dans le cas d'un échantillon prélevé avec probabilités de sélection approximativement égales, la méthode de pondération égale des ménages devrait s'approcher de la solution optimale, au moins pour le cas où les membres du ménage au temps t viennent d'un ou deux ménages de la vague initiale. Ce serait le cas, par exemple, d'un échantillon initial prélevé avec probabilités égales, avec peut-être seulement quelques cas de probabilités inégales. Avec la méthode de pondération égale des ménages, quand seulement un des C_j ménages initiaux, h_j , représentés dans H_i a été sélectionné dans l'échantillon initial (comme ce sera généralement le cas), le poids de H_i est simplement $1/C_j p_j$.

Dans le cas d'un échantillon initial avec probabilités de sélection inégales, le choix des α_i dépendrait idéalement des probabilités de sélection des ménages initiaux. Toutefois, puisque ces probabilités sont inconnues, on ne peut utiliser cette approche. En remplacement, les méthodes de pondération égale des ménages et des personnes peuvent être employées. L'application de ces méthodes (ou de toute méthode fondée sur des constantes α_{ij} respectant la condition $\sum_j \alpha_{ij} = 1$) à un échantillon initial avec probabilités de sélection inégales donne quand même une estimation sans biais \bar{Y} . Le seul inconvénient de ces méthodes dans un tel cas vient du fait que les α_{ij} sont sous-optimaux pour ce qui est de minimiser la variance de \bar{Y} .

Notons que la méthode de pondération égale des ménages exige la connaissance du nombre de ménages initiaux h_j fournissant des membres au ménage H_i au temps t . Ce nombre peut parfois être difficile à établir. Prenons le cas, par exemple, d'un ménage qui, au temps t , comprend deux cohabitants. Il pourrait être difficile de déterminer si ces deux personnes appartenaient au même ménage ou à des ménages différents au moment de la sélection de l'échantillon initial. La méthode de pondération égale des personnes est intéressante du fait qu'elle n'exige pas d'information sur les ménages de la première vague, sauf pour les personnes importantes qui justifient l'emploi de la méthode de pondération égale des personnes de préférence à la méthode de pondération égale des ménages.

4. ESTIMATIONS TRANSVERSALES AU NIVEAU DES PERSONNES

Lorsqu'on produit des estimations transversales au niveau des personnes pour toute vague d'un panel de ménages au-delà de la première, il faut tenir compte du fait qu'un certain nombre de nouveaux venus se seront joints à la population de l'enquête depuis le début du panel. Les

comme la méthode de pondération fondée sur des parts équitables (Huang 1984; Ernst 1989). Ici, nous l'appellerons *méthode de pondération égale des personnes*. Selon cette

méthode,

$$w_i = \frac{1}{\sum_j M_j} M_i w_j^*,$$

où $w_j^* = w_j^k$ est constant pour toutes les personnes du ménage H_i émanant du même ménage inclus dans l'échantillon au temps 1, M_j est le nombre de personnes du ménage H_i venant du ménage H_j , et $M_i = \sum_j M_j$ est le nombre de personnes du ménage H_i qui étaient admissibles à la sélection au temps 1. La méthode de pondération égale des personnes est appliquée dans l'enquête SIPP et on envisage de l'implanter dans l'EDTR.

On voit tout de suite que cette méthode, bien qu'elle soit décrite ici en termes de personnes plutôt que de ménages, pourrait tout aussi bien être décrite en termes de ménages. Comme on l'a vu plus haut, le poids du ménage $w_i = \sum_j \alpha_{ij} w_j^*$ répond à la condition $E(w_i) = 1$ pour n'importe quel ensemble de constantes α_{ij} tel que $\sum_j \alpha_{ij} = 1$. Dans la méthode de pondération égale des ménages, on choisit $\alpha_{ij} = 1/C_i$, avec $\sum_j \alpha_{ij} = 1$. Le choix $\alpha_{ij} = M_j/M_i$, avec $\sum_j \alpha_{ij} = 1$, donne la méthode de pondération égale des personnes.

Il est instructif de comparer, dans un cas simple, la méthode de pondération avec probabilités de sélection inverses et les méthodes de pondération égale des ménages et des personnes. À l'exemple de Little (1989), prenons le cas d'un ménage H_i prélevé au temps t dont les membres viennent de deux ménages initiaux. Soient p_1 et p_2 les probabilités de sélection des ménages initiaux, et p_{12} leur probabilité de sélection conjointe. Selon la méthode des probabilités de sélection inverses, le poids du ménage est

$$w_i^* = \frac{p_1 + p_2 - p_{12}}{1},$$

comme il a été indiqué plus haut.

Selon la méthode de pondération égale des personnes, le poids du ménage H_i dépend du ou des ménages qui ont été prélevés dans l'échantillon initial:

$w_i = p_1/p_1$ si le ménage H_i seul a été sélectionné; $w_i = p_2/p_2$ si le ménage H_2 seul a été sélectionné; $w_i = (p_1/p_1) + (p_2/p_2)$ si H_1 et H_2 ont tous deux été sélectionnés;

où p_1 et p_2 sont les proportions des membres du ménage H_i qui venaient des ménages H_1 et H_2 respectivement (en excluant les nouveaux venus dans la population). La probabilité que seul le ménage H_1 soit sélectionné est $(p_1 - p_{12})$, que seul le ménage H_2 soit sélectionné est $(p_2 - p_{12})$, et que les deux ménages soient sélectionnés est p_{12} . L'espérance du poids, à la condition que le ménage H_i soit dans l'échantillon, est donc

$$E(w_i | H_i \text{ dans l'éch.}) =$$

$$\frac{(p_1 - p_{12})(p_1/p_1) + (p_2 - p_{12})(p_2/p_2) + p_{12}[(p_1/p_1) + (p_2/p_2)]}{p_1 + p_2 - p_{12}},$$

c'est-à-dire,

$$E(w_i | H_i \text{ dans l'éch.}) = \frac{p_1 + p_2 - p_{12}}{1} = w_i^*.$$

Comme le démontre ce résultat, le poids du ménage H_i varie selon les ménages initiaux qui ont été sélectionnés, mais en termes d'espérance, il est le même que celui obtenu par la méthode des probabilités de sélection inverses. En ce qui touche l'espérance du poids du ménage H_i en vertu de la méthode de pondération égale des ménages, elle peut être directement obtenue comme cas spécial du raisonnement ci-dessus, avec $p_1 = p_2 = 1/2$. En termes d'espérance, le poids est le même que celui obtenu par la méthode des probabilités de sélection inverses. Comme le poids $w_i = \sum_j \alpha_{ij} w_j^*$ répond à la condition $E(w_i) = 1$ pour n'importe quel ensemble de α_{ij} tel que $\sum_j \alpha_{ij} = 1$, il convient de se demander quel est le choix optimal pour les α_{ij} . Une possibilité consiste à choisir les α_{ij} de façon à minimiser la variance du total estimé \bar{Y} .

La variance de \bar{Y} peut s'écrire

$$V(\bar{Y}) = VE(\bar{Y} | s) + EV(\bar{Y} | s), \quad (3.7)$$

où s désigne l'ensemble des ménages de l'échantillon au temps t . On a alors

$$E(\bar{Y} | s) = E\left(\sum_{H_i} w_i Y_i | s\right) =$$

$$= \sum_s E(w_i | H_i) Y_i = \sum_s w_i^* Y_i = \bar{Y}^*,$$

où \bar{Y}^* est l'estimateur habituel avec probabilités de sélection inverses. Par conséquent,

$$VE(\bar{Y} | s) = V(\bar{Y}^*).$$

Le premier terme de l'équation (3.7) est donc la variance de l'estimateur habituel avec probabilités de sélection inverses, et le deuxième terme est la variance additionnelle résultant de l'utilisation des méthodes de pondération de la classe (3.5), $w_i = \sum_j \alpha_{ij} w_j^*$. Les α_{ij} peuvent alors être choisis de façon à minimiser $EV(\bar{Y} | s)$.

Considérons

$$V(\bar{Y} | s) = V\left(\sum_H w_i Y_i | s\right)$$

$$= \sum_s Y_i^2 V(w_i | H_i) +$$

$$\sum_{i \neq i'} \sum Y_i Y_{i'} \text{Cov}(w_i, w_{i'} | H_i, H_{i'}).$$

En comparant les équations (3.1) et (3.2), on peut voir que X est sans biais pour Y dans le cas de n importe quelle méthode de pondération pour laquelle $E(w_i) = 1$ pour tout i . Il y a de nombreuses façons de remplir la condition $E(w_i) = 1$. Nous allons en examiner trois ici. Prenons d'abord le cas d'une *méthode de pondération courante à probabilités de sélection inverses*. La probabilité qu'un ménage soit dans l'échantillon au temps t est la probabilité qu'un ou plusieurs des ménages du temps t dont il a hérité de membres ait fait partie de l'échantillon initial. La probabilité que le ménage H_i soit dans l'échantillon au temps t est donc

$$P(H_i) = P(h_j \cup h_k \cup h_l \cup \dots)$$

$$= \sum p_j - \sum \sum p_{jk} + \sum \sum \sum p_{jkl} - \dots, \quad (3.3)$$

où $P(h_j \cup h_k \cup h_l \cup \dots)$ est la probabilité de sélection de l'union des ménages initiaux h_j, h_k, h_l, \dots , dans l'échantillon initial, p_j est la probabilité de sélection du ménage initial h_j dans l'échantillon initial, p_{jk} est la probabilité de sélection conjointe des ménages initiaux h_j et h_k dans l'échantillon initial, \dots , et où les ménages h_j, h_k, h_l, \dots contiennent chacun au moins un membre qui se trouve maintenant dans le ménage H_i . Le poids de chaque ménage de l'échantillon est alors $w_i = 1/P(H_i)$. Selon cette méthode de pondération,

$$E(w_i) = P(H_i) [1/P(H_i)] + [1 - P(H_i)] 0 = 1,$$

ce qui remplit la condition énoncée pour l'obtention d'un estimateur sans biais d'un total de la population.

En pratique, le calcul de $P(H_i)$ ne sera généralement pas aussi complexe que l'équation (3.3) pourrait le laisser craindre, car le nombre de ménages initiaux représentés dans le ménage H_i est habituellement peu élevé. Avec, disons, deux ménages originaux, $P(H_i)$ se réduit à

$$(3.4) \quad P(H_i) = P(h_1 \cup h_2) = p_1 + p_2 - p_{12}.$$

L'application de la méthode des probabilités de sélection inverses pose un problème, car il se peut que p_j soit connu seulement pour les ménages inclus dans l'échantillon initial et non pour les autres ménages. Il se peut aussi que les probabilités conjointes soient inconnues. Même quand l'échantillon initial a été prélevé avec probabilités égales, de sorte que tous les p_j sont les mêmes, la probabilité conjointe peut dépendre du plan de sondage (elle peut varier, par exemple, selon que les deux ménages appartiennent au même segment ou non). La difficulté d'obtenir $P(H_i)$ est un important inconvénient de la méthode des probabilités de sélection inverses.

Une autre stratégie permettant de déterminer les poids bilités de sélection des ménages inclus dans l'échantillon initial, ce qui évite le problème de l'obtention de $P(H_i)$ comme ci-dessus. On peut le faire, par exemple, en déterminant l'ensemble de ménages h_j au temps t expliquant la présence du ménage H_i dans l'échantillon au temps t , et en calculant le poids du ménage H_i de la façon suivante:

$$E(w_{ijk}^*) = p_j(1/p_j) + (1 - p_j)0 = 1.$$

Dans ce cas, le choix naturel pour les constantes α_{ijk} est de les considérer égales pour tous les membres du ménage courant qui étaient admissibles à la sélection dans l'échantillon initial. On obtient ainsi ce qui a été désigné

$$w_i = \sum_k \sum_j \alpha_{ijk} w_{ijk}^*,$$

où $w_{ijk}^* = 1/p_j$ si la personne k du ménage h_j était dans l'échantillon initial et $w_{ijk}^* = 0$ sinon, et où les α_{ijk} sont n'importe quel ensemble de constantes répondant à la condition $\sum_j \sum_k \alpha_{ijk} w_{ijk}^* = 1$. Puisque la probabilité qu'une personne soit incluse dans l'échantillon initial est la même que celle du ménage auquel elle appartient, on a

$$(3.6) \quad w_i = \sum_j w_{ij}^*/C_j,$$

où C_j est le nombre de ménages initiaux représentés dans le ménage H_i au temps t . Une autre version de cette même approche se fonde sur les personnes, plutôt que sur les ménages, faisant partie de l'échantillon initial. Dans ce cas, supposons que I_{ijk} désigne la personne k du ménage initial j dans le ménage i . On a alors

$$E(w_i) = \sum_j \alpha_{ij} = 1.$$

et donc

$$E(w_{ij}^*) = p_j(1/p_j) + (1 - p_j)0 = 1,$$

Selon cette approche,

de constantes répondant à la condition $\sum_j \alpha_{ij} = 1$.

où $w_{ij}^* = 1/p_j$ si le ménage h_j , qui a au moins un membre dans le ménage H_i , faisait partie de l'échantillon initial et $w_{ij}^* = 0$ sinon, et où les α_{ij} sont n'importe quel ensemble

$$(3.5) \quad w_i = \sum_j \alpha_{ij} w_{ij}^*,$$

ménage est trois fois plus élevée. L'analyse des données résultantes exige par conséquent de recourir à des méthodes de pondération qui compensent ces probabilités de sélection inégales.

La composition de la population change quand des personnes entrent dans la population ou en sortent. Une personne membre de l'échantillon au temps 1 qui sort de la population avant le temps 2 réduit la taille de l'échantillon du temps 2, mais n'influe pas par ailleurs sur les estimations transversales du temps 2. Essentiellement, la base d'échantillonnage pour la population du temps 2 est la population du temps 1, et les personnes qui sont sorties dans l'intervalle sont traitées comme des blancs dans la base. La simple omission des blancs sélectionnés dans l'échantillon du temps 2 ne cause pas de biais aux estimations de l'enquête (voir, par exemple, Kish 1965). Toutefois, le cas des nouveaux venus, c'est-à-dire des personnes qui entrent dans la population, n'est pas aussi simple. En vertu de la règle de dénombrement des enquêtes à panel de ménages décrite ci-dessus, les nouveaux venus qui se joignent à des ménages comprenant des personnes qui étaient admissibles à l'échantillon initial sont incorporés à la population en vue des estimations transversales faites à des moments ultérieurs. Toutefois, les nouveaux venus qui créent leurs propres ménages ne sont pas représentés dans les analyses faites au niveau des personnes aux vagues ultérieures du panel. De même, les ménages composés uniquement de nouveaux venus ne sont pas représentés dans les analyses faites au niveau des ménages lors des vagues ultérieures.

L'omission des enquêtes par panel menées auprès des ménages d'inclure les ménages formés uniquement de nouveaux venus présente un problème pour les analyses transversales aux vagues ultérieures du panel. Si ces ménages et leurs membres constituent une proportion négligeable de la population, la solution peut être simplement de ne pas en tenir compte. Toutefois, si leur proportion est appréciable, ce qui peut arriver aux dernières vagues d'un panel de longue durée, des solutions de rechange peuvent être utilisées. Une possibilité consiste à ajouter au panel un échantillon formé de nouveaux venus (p. ex. des immigrants), comme le décrit Lavalée (1995) au sujet de l'EDTR. Il arrive souvent, cependant, que cette solution ne soit pas applicable en pratique. Une autre solution consiste à limiter la population d'inférence aux personnes qui étaient membres de la population au début du panel. Les nouveaux venus qu'on trouve dans des ménages de membres de l'échantillon sont alors exclus de l'échantillon. Cette solution offre une délimitation claire de la population d'inférence. Elle conviendra dans la mesure où une telle définition répond adéquatement aux objectifs de l'enquête.

Les changements de la composition de la population posent des problèmes pour les analyses longitudinales au niveau des personnes. Dans bien des applications, la population d'inférence est limitée aux personnes qui étaient présentes dans la population pendant toute la période d'observation visée par l'analyse. L'inclusion de cohabitants dans l'analyse longitudinale est aussi une cause de problèmes. Si la période visée par l'analyse longitudinale

commence au début du panel, l'analyse peut simplement être limitée aux membres de l'échantillon initial. Si la période commence plus tard, il est tentant d'inclure à la fois les membres de l'échantillon initial et les cohabitants qui se sont joints au panel avant le début de la période visée par l'analyse. Toutefois, les règles de dénombrement habituelles des enquêtes par panel menées auprès des ménages précisent que les cohabitants sont suivis tant qu'ils continuent à vivre avec des membres de l'échantillon initial, mais qu'on cesse de recueillir des données à leur sujet s'ils ne vivent plus avec ces personnes. À moins que la période d'analyse soit assez courte pour que le nombre de cohabitants qui cessent de vivre avec des personnes de l'échantillon pendant cette période soit négligeable, cette règle de dénombrement rend problématique l'inclusion des cohabitants dans les analyses longitudinales. Ce problème est examiné plus en détail à la section 5.

3. ESTIMATIONS TRANSVERSALES AU NIVEAU DES MÉNAGES

Dans la présente section, nous examinons des méthodes de pondération qui permettent de produire des estimations transversales au niveau des ménages pour n'importe quelle vague, outre la première, d'une enquête par panel menée auprès des ménages. À la première vague, un échantillon de ménages est prélevé et toutes les personnes des ménages de l'échantillon deviennent les membres du panel, qui seront suivies jusqu'à la fin du panel ou jusqu'à ce qu'elles sortent de la population de l'enquête. À une vague subséquente, la vague t , l'échantillon de ménages comprend tous les ménages dans lesquels vivent des membres du panel. Les ménages formés seulement de nouveaux venus ne sont pas représentés dans l'échantillon aux vagues subséquentes. Ces ménages ne sont pas pris en considération ici. Les complications liées à la non-réponse seront examinées à la section 6.

Considérons l'estimation du total Y pour l'ensemble des H ménages de la population au temps t :

$$Y = \sum_{h=1}^H Y_h \tag{3.1}$$

Un estimateur général pour ce total peut s'écrire de la façon suivante:

$$Y = \sum_{h=1}^H w_h Y_h$$

où w_h est une variable aléatoire qui prend la valeur $w_h = 0$ si le ménage h n'est pas dans l'échantillon. L'espérance de Y est

$$E(Y) = \sum_{h=1}^H E(w_h) Y_h \tag{3.2}$$

2. ÉVOLUTION DE LA COMPOSITION DE LA POPULATION ET DES MÉNAGES DANS LE TEMPS

non-réponse et du sous-dénombrement. Notre analyse s'appuie largement sur des travaux antérieurs de Ernst (1989), Gailly et Lavallée (1993), Huang (1984), Judkins et coll. (1984), Lavallée et Hunter (1992) et Little (1989). La section 6 examine ensuite brièvement la question des rajustements apportés aux poids pour compenser les données manquantes dues à la non-réponse et au sous-dénombrement. La section 7, enfin, énonce certaines conclusions et présente un autre exemple auquel peut s'appliquer la méthode du partage des poids.

Dans l'analyse d'une enquête par panel, il faut tenir compte du fait que la population de l'enquête change avec le temps. Dans le cas des enquêtes par panel menées auprès des ménages, il importe de faire la distinction entre l'évolution de la composition de la population et celle de la composition des ménages.

La composition de la population d'une enquête change avec le temps parce que certaines personnes sortent de la population, d'autres y entrent, et certaines peuvent sortir de la population et y entrer plus d'une fois. Les décès, l'émigration ou l'entrée en institution (pour les enquêtes qui n'incluent pas les personnes en institution) sont des causes de sortie de la population. La composition change aussi quand des membres entrent dans la population au moment de la naissance (ou lorsqu'ils atteignent l'âge minimum nécessaire), ou encore lorsqu'ils immigreront ou qu'ils sortent d'une institution.

La composition des ménages change avec le temps pour de nombreuses raisons, comme les décès, les mariages et les divorces. Par exemple, un ménage existant au temps 1 peut comprendre plusieurs personnes qui aboutiront dans plusieurs ménages différents au temps 2. Ces personnes peuvent créer de nouveaux ménages, se joindre à des personnes qui appartenaient à un ou plusieurs ménages au temps 1, ou encore se joindre à des personnes qui n'étaient pas dans la population au temps 1. Certaines personnes peuvent aussi sortir de la population dans l'intervalle.

Prenons le cas d'un plan d'échantillonnage simple dans lequel les ménages sont prélevés de façon indépendante, avec probabilités égales, au temps 1. Au temps 2, l'échantillon de ménages comprend tous les ménages qui incluent une ou plusieurs personnes qui appartenaient aux ménages de l'échantillon prélevé au temps 1, et l'échantillon de personnes comprend tous les membres des ménages qui font partie de l'échantillon au temps 2. Les échantillons de ménages et de personnes au temps 2 sont prélevés avec probabilités inégales. Par exemple, la probabilité de sélection au temps 2 d'un ménage comprenant des personnes qui viennent de trois ménages qui existaient au temps 1 est trois fois plus grande que la probabilité de sélection au temps 2 d'un ménage comprenant des personnes qui viennent d'un seul ménage qui existait au temps 1. De même, la probabilité de sélection des personnes de ce

sans biais) des paramètres de la population. Il est utile, dans le contexte des enquêtes par panel menées auprès des ménages, de faire la distinction entre trois types d'analyses :

- analyses transversales des ménages à un point du temps particulier;
- analyses transversales des personnes à un point du temps particulier;
- analyses longitudinales des personnes sur une certaine période.

Des méthodes de pondération s'appliquant à ces trois types d'analyses seront examinées dans les sections suivantes. Les analyses longitudinales des ménages sur une certaine période ne sont pas examinées ici, en raison de la nature problématique de ce type d'analyse, due au fait que la composition des ménages évolue avec le temps (voir, par exemple, Duncan et Hill 1985).

Les méthodes de pondération utilisées dans les enquêtes par panel menées auprès des ménages doivent tenir compte du fait qu'il existe parfois plusieurs voies par lesquelles les ménages et les personnes observés à une vague particulière peuvent être sélectionnés. À une vague donnée, un ménage et ses membres sont inclus dans l'échantillon si n'importe quel des ménages initiaux (c.-à-d. les ménages qui existaient au moment de la sélection initiale) ayant fourni des membres au ménage courant faisait partie de l'échantillon initial. Selon la méthode de pondération habituelle, on attribue aux ménages des poids inversement proportionnels à leur probabilité de sélection conjointe, en tenant compte des différentes façons dont ils peuvent être sélectionnés. Toutefois, cette méthode ne convient pas dans la plupart des enquêtes par panel menées auprès des ménages, car ces probabilités de sélection conjointes ne peuvent être déterminées. L'autre méthode de pondération examinée ici, appelée par Lavallée (1995) *méthode du partage des poids*, n'exige pas la connaissance des probabilités de sélection conjointe des éléments de l'échantillon, mais elle introduit une variation aléatoire dans la pondération. Puisque cette variation aléatoire entraîne une réduction de la précision des estimations de l'enquête par rapport à ce qu'on obtiendrait avec une pondération selon les probabilités de sélection inverses, cette méthode substitue ne doit être envisagée que dans des situations où il est impossible de déterminer avec certitude les probabilités de sélection conjointes. Ce genre de situation caractérise souvent les enquêtes par panel menées auprès des ménages, ainsi que certains autres plans de sondage dans lesquels il existe plusieurs voies pour la sélection des éléments.

En guise d'introduction à notre examen des méthodes de pondération s'appliquant aux enquêtes par panel menées auprès des ménages, nous examinons de façon plus détaillée dans la section suivante les changements qui peuvent toucher les ménages au fil du temps, ainsi que les types de personnes concernées. Les sections 3, 4 et 5 examinent ensuite les méthodes de pondération qui peuvent servir aux trois formes d'analyse différentes mentionnées ci-dessus. Ces sections traitent de méthodes de pondération pour des probabilités de sélection inégales, sans les compli-

Méthodes de pondération pour les enquêtes par panel auprès des ménages

GRAHAM KALTON et J. MICHAEL BRICK¹

RÉSUMÉ

Il est fréquent, dans les enquêtes par panel menées auprès des ménages, de prélever d'abord un échantillon de ménages et de tenter par la suite de suivre tous les membres de ces ménages pendant la durée du panel. Aux vagues subséquentes, les données sont recueillies pour les membres de l'échantillon initial et pour toutes les personnes qui vivent avec les membres de l'échantillon à ce moment. Il est souhaitable, en effet, d'inclure les données recueillies aussi bien auprès des membres de l'échantillon initial que des personnes qui vivent avec elles lorsqu'on fait des estimations transversales au niveau des personnes pour une vague particulière. De même, il est souhaitable d'inclure les données de tous les ménages pour lesquels des données sont recueillies à une vague particulière lorsqu'on fait des estimations transversales au niveau des ménages pour cette vague. Le présent article examine des méthodes de pondération qui peuvent être utilisées à cette fin. Ces méthodes de pondération peuvent aussi servir dans d'autres contextes, lorsqu'il existe plusieurs voies par lesquelles les unités peuvent être prélevées dans l'échantillon.

MOTS CLÉS: Estimations transversales; pondération fondée sur des parts équitables; pondération de multiplicité; enquêtes par panel; méthode du partage des poids.

1. INTRODUCTION

De nombreux pays ont mis sur pied, ces dernières années, des enquêtes par panel destinées à étudier l'économie des ménages. L'enquête américaine PSID (Panel Study of Income Dynamics), menée par le Survey Research Center de l'université du Michigan, a été lancée en 1968 et a permis de recueillir des données annuellement depuis ce temps (Hill 1992), tandis que la British Household Panel Survey a été amorcée en 1990 (Buck et coll. 1994). Des enquêtes par panel du même genre sont en cours ou en voie d'être implantées dans la plupart des autres pays européens. Le US Bureau of the Census a mis en marche l'enquête SIPP (Survey of Income and Program Participation) en 1983 (Nelson et coll. 1985; Kasprzyk 1988; Jabine et coll. 1990; Citro et Kalton 1993), et Statistique Canada a lancé l'Enquête sur la dynamique du travail et du revenu (EDTR) en 1994 (Lavallée et coll. 1993).

Dans la plupart de ces enquêtes par panel menées auprès des ménages, on commence par prélever un échantillon national de ménages, et l'on suit tous les membres de ces ménages pendant la durée du panel. Au fil du temps, divers changements sont susceptibles de modifier la composition des ménages. Certains membres des ménages de l'échantillon initial s'en vont former leur propre ménage ou se joignent à un autre, par exemple lorsqu'une fille quitte le foyer familial pour se marier. De nouveaux membres peuvent se joindre aux ménages de l'échantillon initial, par exemple lorsqu'un parent âgé vient habiter chez un de ses enfants ou qu'un nouveau-né vient grossir les rangs de la famille. Afin de pouvoir décrire la situation économique des membres de l'échantillon à différents moments, les responsables des enquêtes par panel auprès des ménages recueillent habituellement des données non seulement sur

les membres de l'échantillon, mais aussi sur les personnes qui vivent avec eux au moment où les observations sont faites. Comme Lavallée (1995), nous appelons ces personnes cohabitantes dans le présent article. Dans d'autres documents, on parle souvent de personnes associées ou de personnes extérieures à l'échantillon.

Avec l'augmentation de la durée du panel, la proportion des cohabitantes qui font partie de l'échantillon aux différentes vagues s'accroît. Par exemple, pour le panel de 1984 de l'enquête SIPP, les cohabitantes représentaient environ 8,6% de l'échantillon après un an et environ 12,6% de l'échantillon après deux ans (d'après le tableau 1 de Kasprzyk et McMillen 1987). Lorsque les enquêtes ont des panels de longue durée, la proportion des cohabitantes devient considérable après quelques années. Selon la définition de l'enquête PSID, par exemple, les membres de l'échantillon sont l'ensemble des membres des unités familiales échantillonnées en 1968 qui sont toujours vivants, l'ensemble des enfants nés des membres de l'échantillon initial depuis le début du panel, et les enfants de ces enfants. Des données sont aussi recueillies, dans l'enquête PSID, sur les cohabitantes qui vivent avec les membres de l'échantillon à chaque vague de collecte des données. Des l'échantillon initial nés depuis le début du panel ou des l'échantillon initial, 34,6% étaient des membres de l'échantillon initial, 41,2% étaient des membres de l'échantillon initial, et 24,2% étaient des cohabitantes (ces chiffres ne tiennent pas compte de l'échantillon latino-américain ajouté en 1990) (Hill 1995).

Dans le présent article, nous examinons des méthodes de pondération des données recueillies aussi bien auprès des membres de l'échantillon que des cohabitantes, qui visent à produire des estimations sans biais (ou approximativement

¹ Graham Kalton et J. Michael Brick, Westat Inc., 1650 Research Blvd., Rockville, MD 20850, U.S.A.

SINGH, M.P., DREW, J.D., GAMBINO, J.G., et MAYDA, F. (1990), *Méthodologie de l'enquête sur la population active du Canada*, Statistique Canada, n° 71-526 au catalogue. THOMPSON, S.K. (1992), *Sampling*, New York: John Wiley and Sons.

WOLTER, K.M. (1985), *Introduction to Variance Estimation*. New York: Springer-Verlag.
YATES, F., et GRUNDY, P.M. (1953), Selection without replacement from within strata with probability proportional to size, *Journal of the Royal Statistical Society B*, 15, 235-261.

puisque en sélectionnant des ménages complets, $\pi_{(1)}^{Ij} = \pi_{(1)}^{Ij}$ pour $j \in I$. La variance $\text{Var}(Z_{*(1)}^*)$ s'obtient alors directement par

$$\text{Var}(Z_{*(1)}^*) = \sum_{i=1}^{N_{(1)}} \sum_{I'=1}^I \frac{\pi_{(1)}^{I'} \pi_{(1)}^{I'}}{(\pi_{(1)}^{I'} - \pi_{(1)}^{I'} \pi_{(1)}^{I'})} Z_{*(1)}^{I'} Z_{*(1)}^{I'}. \quad (13)$$

En ce qui a trait à $Z_{*(2)}^*$, les individus peuvent également être indexés à nouveau aux fins de l'harmonisation avec $Z_{*(1)}^*$, même si cette modification est sans effet sur la forme de $Z_{*(2)}^*$. En suivant la voie utilisée pour $\text{Var}(Z_{*(1)}^*)$, on obtient $\text{Var}(Z_{*(2)}^*)$ avec la formule suivante:

$$\text{Var}(Z_{*(2)}^*) = \sum_{i=1}^{N_{*(2)}} \sum_{I'=1}^I \frac{\pi_{*(2)}^{I'} \pi_{*(2)}^{I'}}{(\pi_{*(2)}^{I'} - \pi_{*(2)}^{I'} \pi_{*(2)}^{I'})} Z_{*(2)}^{I'} Z_{*(2)}^{I'}, \quad (14)$$

où $N_{*(2)}$ représente le nombre de ménages de l'année 2 contenant au moins un immigrant et $Z_{*(2)}^{I'} = \sum_{j=1}^{M_{*(2)}^{I'}} z_{ij}^{I'}$. La quantité $M_{*(2)}^{I'}$ représente le nombre d'immigrants présents dans le ménage I' .

Finalement, $\text{Var}(Y)$ s'obtient simplement par

$$\text{Var}(Y) = \sum_{i=1}^{N_{(1)}} \sum_{I'=1}^I \frac{\pi_{(1)}^{I'} \pi_{(1)}^{I'}}{(\pi_{(1)}^{I'} - \pi_{(1)}^{I'} \pi_{(1)}^{I'})} Z_{*(1)}^{I'} Z_{*(1)}^{I'}$$

$$+ \sum_{i=1}^{N_{*(2)}} \sum_{I'=1}^I \frac{\pi_{*(2)}^{I'} \pi_{*(2)}^{I'}}{(\pi_{*(2)}^{I'} - \pi_{*(2)}^{I'} \pi_{*(2)}^{I'})} Z_{*(2)}^{I'} Z_{*(2)}^{I'}. \quad (15)$$

La variance (15) peut être estimée sans biais à l'aide de l'équation suivante:

$$\widehat{\text{Var}}(Y^*) = \sum_{i=1}^{n_{(1)}} \sum_{I'=1}^I \frac{\pi_{(1)}^{I'} \pi_{(1)}^{I'}}{(\pi_{(1)}^{I'} - \pi_{(1)}^{I'} \pi_{(1)}^{I'})} Z_{*(1)}^{I'} Z_{*(1)}^{I'}$$

$$+ \sum_{i=1}^{n_{*(2)}} \sum_{I'=1}^I \frac{\pi_{*(2)}^{I'} \pi_{*(2)}^{I'}}{(\pi_{*(2)}^{I'} - \pi_{*(2)}^{I'} \pi_{*(2)}^{I'})} Z_{*(2)}^{I'} Z_{*(2)}^{I'}. \quad (16)$$

Comme l'EDTR est en fait un sous-échantillon de l'EPA, l'estimateur de variance jackknife élaboré pour l'EPA (voir Singh et coll., 1990) peut également servir ici, après quelques légères modifications. La méthode jackknife fonctionne habituellement comme suit: l'échantillon est d'abord divisé en groupes aléatoires (ou en répliques, selon la terminologie de l'EPA). Ensuite, chaque groupe aléatoire r est retiré à tour de rôle de l'échantillon, et une nouvelle estimation $Y^{(r)}$ du Y total est calculée. Les estimations différentes $Y^{(r)}$ sont finalement comparées à l'estimation originale Y pour obtenir une estimation de la variance $\text{Var}(Y)$. Pour de plus amples détails sur la méthode jackknife en général, voir Särndal, Swensson et Wretman (1992).

BIBLIOGRAPHIE

ERNST, L. (1989). Weighing issues for longitudinal household and family estimates. Dans *Panel Surveys*. (Éds. D. Kasprzyk, G. Duncan, G. Kalton et M.P. Singh). New York: John Wiley and Sons, 135-159.

GAILLY, B., et LAVALLÉE, P. (1993). Insérer des nouveaux membres dans un panel longitudinal de ménages et d'individus: simulations. CEPS/Instead, Document PSELL No. 54, Luxembourg, mai 1993.

HORVITZ, D.G., et THOMPSON, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.

KALTON, G., et BRICK, J.M. (1995). Méthodes de pondération pour les enquêtes par panel auprès des ménages. *Techniques d'enquête*, 21, 37-49.

LAVALLÉE, P. (1993). Représentativité de l'échantillon de l'Enquête sur la dynamique du travail et du revenu. Statistique Canada, document de recherche de l'Enquête sur la dynamique du travail et du revenu, n° 93-19 au catalogue, Décembre 1993.

LEMAÎTRE, G., et DUFOUR, J. (1987). Une méthode intégrée de pondération des personnes et des familles. *Techniques d'enquête*, 13, 211-220.

SÄRNDAL, C.-E., SWENSSON, B., et WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

REMERCIEMENTS

L'auteur souhaite remercier MM. Carl Särndal et M.P. Singh, ainsi que le rédacteur associé et les lecteurs dont les commentaires extrêmement utiles lui ont permis de clarifier le texte du présent article. Il remercie aussi sincèrement M. Jean-Claude Deville, qui a suggéré d'étendre la méthode du partage des poids à un contexte général.

la présente discussion.

$$\widehat{\text{Var}}_j(Y) = \sum_h \frac{R_h}{(R_h - 1)} \sum_{r \in h} (Y^{(hr)} - Y)_2, \quad (17)$$

suivante:

Il convient de rappeler que l'EPA s'appuie sur un plan stratifié à plusieurs degrés qui utilise une base de sondage aréolaire. À l'intérieur de chaque strate h de premier degré, les groupes aléatoires (ou les répliques) correspondent essentiellement aux unités primaires d'échantillonnage (UPÉ). Pour calculer l'estimation de variance jackknife pour l'estimation du Y total, on peut utiliser la formule

3.2.2 Calcul des poids de base

Nous appliquerons maintenant la méthode de partage des poids décrite à la section 2 à l'échantillon de l'EDTR et à l'échantillon supplémentaire. La population U^A représente ici la réunion de deux populations distinctes $U^{(1)}$ et $U^{(2)}$, soit $U^A = U^{(1)} + U^{(2)}$. L'échantillon s^A de $m = m^{(1)} + m^{(2)}$ individus correspond à la réunion des deux échantillons distincts $s^{(1)}$ et $s^{(2)}$. La population U^B est représentée par $U = U^{(1)} + U^{(2)}$. La population $U^A = U^*$ exclut les nouveaux-nés tandis que la population $U^B = U$ les inclut. Les grappes de populations U^B correspondent simplement aux N ménages de l'année 2, ce qui fait que $M_B^i = M_i^i$.

Un lien possible peut être établi entre les populations U^A et U^B par les individus qui font partie des deux populations. Ainsi, $l_{jk} = 1$ si l'individu j de la population U^A correspond à l'individu k de la population U^B , et $l_{jk} = 0$ dans les autres cas. Pour chaque individu k qui n'est pas un nouveau-né, nous obtenons alors $L_{ik} = \sum_{j=1}^{M_A^i} l_{j,ik} = 1$. Par contre, pour chaque nouveau-né k , nous obtenons $L_{ik} = \sum_{j=1}^{M_A^i} l_{j,ik} = 0$ puisque les nouveaux-nés sont exclus de U^A . Nous obtenons ainsi $L_i = \sum_{k=1}^{M_B^i} L_{ik} = M_i^i$ où M_i^i représente la taille du ménage i à l'exclusion des nouveaux-nés.

Souignons qu'il s'agit là d'un type de lien parmi de nombreuses possibilités. On pourrait ainsi également étendre le lien décrit au paragraphe précédent à tous les autres membres du ménage, c'est-à-dire en déterminant que $l_{jk} = 1$ pour tous les individus k (de U^B) appartenant au même ménage i où l'individu j (de U^A) appartient maintenant à U^B , et 0 dans les autres cas. En d'autres mots, $l_{jk} = 1$ si les individus j et k appartiennent au même ménage i . Pour chaque individu k du ménage i , nous obtenons ainsi $L_{ik} = \sum_{j=1}^{M_A^i} l_{j,ik} = M_i^i$. Nous obtenons également $L_i = \sum_{k=1}^{M_B^i} L_{ik} = \sum_{k=1}^{M_B^i} M_i^i = M_B^i M_i^i$. On peut montrer que ce lien donne la même pondération de base que celui décrit au paragraphe précédent. Comme le premier lien correspond à une façon plus naturelle de lier les individus (c.-à-d., liaison limitée aux individus qui sont les mêmes dans les populations U^A et U^B), c'est celle que nous retiendrons ci-après.

À partir de la définition (2) du poids initial w_{ik}^i de l'individu k du ménage i , nous obtenons

$$w_{ik}^i = \frac{f_{ik}^{(1)}}{f_{ik}^{*(1)}} + \frac{\pi_{ik}^{(2)}}{f_{ik}^{*(2)}} \quad (8)$$

où $f_{ik}^{(1)} = 1$ si l'individu k fait partie de $s^{(1)}$, et 0 dans les autres cas, et où $f_{ik}^{*(2)} = 1$ si l'individu k fait partie de $s^{*(2)}$, et 0 dans les autres cas. On peut reformuler le tout plus explicitement comme suit:

$$w_{ik}^i = \begin{cases} 1 / \pi_{ik}^{(1)} & \text{pour } k \in s^{(1)} \\ 1 / \pi_{ik}^{*(2)} & \text{pour } k \in s^{*(2)} \\ 0 & \text{autrement.} \end{cases} \quad (9)$$

On obtient ainsi

$$Z_{*(1)} = \sum_{k=1}^m \frac{\pi_{ik}^{(1)}}{Z_{ik}^*} = \sum_{k=1}^m \sum_{j=1}^n \frac{\pi_{ij}^{(1)}}{Z_{ij}^*} \quad (10)$$

À noter que la première ligne de (9) correspond aux individus longitudinaux. La deuxième ligne correspond aux immigrants choisis par l'intermédiaire de l'échantillon supplémentaire. La troisième ligne représente l'ensemble des nouveaux-nés, des cohabitants (si le ménage est un ménage longitudinal ne faisant pas partie de l'échantillon supplémentaire) ou des individus présents au départ (si le ménage fait partie de l'échantillon supplémentaire).

Le poids de base w_i du ménage i s'obtient à partir de

$$w_i = \frac{\sum_{k=1}^{M_i^i} w_{ik}^i}{\sum_{k=1}^{M_i^i} L_{ik}} = \frac{1}{M_i^i} \sum_{k=1}^{M_i^i} w_{ik}^i, \quad (11)$$

Comme nous l'avons montré à la section 2, cet estimateur est non biaisé pour Y .

3.2.3 Estimation de la variance

La formule de la variance pour Y est fournie par l'équation (6). Toutefois, en présument que les deux échantillons $s^{(1)}$ et $s^{*(2)}$ sont choisis indépendamment, nous obtenons $\text{Var}(Y) = \text{Var}(Z_{*(1)}) + \text{Var}(Z_{*(2)})$, où chaque terme prend la forme de l'équation (6). Pour l'EDTR, cette supposition d'indépendance tient si la sélection de l'échantillon supplémentaire passe par l'EPA.

En ce qui a trait à $Z_{*(1)}$, nous pouvons indexer à nouveau les individus pour refléter le fait que $m^{(1)}$ individus ont été sélectionnés à l'année 1 dans $n^{(1)}$ ménages. On obtient ainsi

$$Z_{*(1)} = \sum_{k=1}^m \frac{\pi_{ik}^{(1)}}{Z_{ik}^*} = \sum_{k=1}^n \sum_{j=1}^n \frac{\pi_{ij}^{(1)}}{Z_{ij}^*} \quad (12)$$

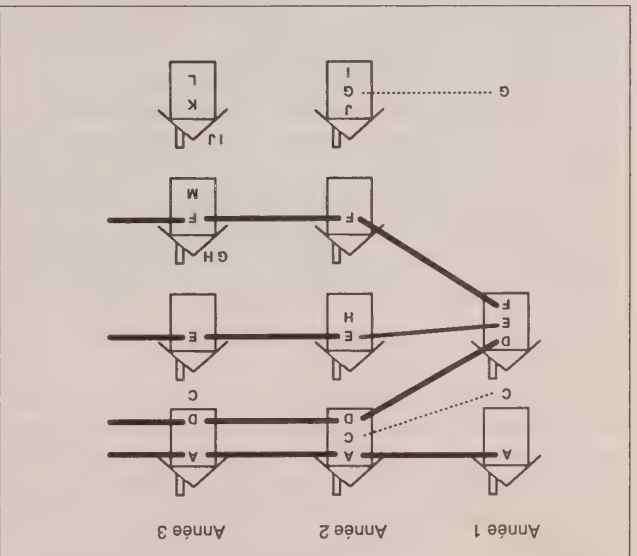


Figure 2. Sélection des individus pour l'EDTR.

Les individus absents ou présents au départ qui se joignent à un ménage longitudinal sont appelés "cohabitants". L'individu H de l'année 2 est un individu absent au départ qui se joint à l'échantillon à titre de cohabitant. La quatrième maison de l'année 2 représente un ménage sélectionné pour l'échantillon supplémentaire de l'année 2 et dans lequel les individus I et J sont des individus absents au départ (l'un d'eux étant nécessairement un immigrant puisque l'échantillon supplémentaire est limité à cette catégorie). L'individu G est un individu présent au départ qui a le même statut que l'individu C. À l'année 3, les individus C et H ont quitté leurs ménages longitudinaux et ne seront donc pas interviewés. Les individus I et J qui ont été inclus dans l'échantillon supplémentaire sont maintenant remplacés par les individus de l'échantillon supplémentaire de l'année 3, c'est-à-dire K et L. L'individu M est un individu absent au départ qui se joint au ménage longitudinal à titre de cohabitant. Il convient finalement de souligner qu'aux fins de l'échantillonnage transversal, un ménage sélectionné peut contenir un ou plusieurs individus longitudinaux, individus présents ou individus absents au départ (nouveaux-nés et immigrants).

3.2 Pondération de base

3.2.1 Considérations générales

Pour produire des estimations transversales, il convient de pondérer l'échantillon longitudinal additionné des individus absents et présents au départ. La première étape consiste à obtenir un *poids de base* pour chaque individu dans chaque ménage interviewé. Le poids de base est le poids avant l'ajustement après sélection et la post-stratification. Il s'agit, d'une certaine façon, de l'équivalent du poids de sondage. Noter que les poids de base servent uniquement à l'estimation transversale. Les poids de base sont obtenus à partir des probabilités de sélection. Comme nous l'avons déjà expliqué plus haut,

La sélection des immigrants par l'intermédiaire d'une sélection de ménages a notamment pour conséquence d'inclure avec l'échantillon supplémentaire d'autres individus (nouveaux-nés, individus présents au départ ou individus longitudinaux) du seul fait qu'ils font partie du ménage de l'immigrant sélectionné. Toutefois, comme les unités de sélection de l'échantillon supplémentaire sont limitées aux immigrants, ces autres individus sont sélectionnés d'une manière irrégulière, même s'ils finissent par être interviewés. Les probabilités de sélection de ces individus sont en réalité mal définies.

Pour assurer la représentativité transversale, certains immigrants sont choisis à même l'échantillon supplémentaire. À l'année 2, nous choisissons alors un échantillon $s^{*(2)}$ de $m^{*(2)}$ individus à partir de la population $U^{*(2)}$ de $M^{*(2)}$ immigrants. L'échantillon supplémentaire est tiré des ménages (les $m^{*(2)}$ ménages). Nous désignons par $\pi_j^{*(2)}$ la probabilité de sélection de l'immigrant j et nous présumons que $\pi_j^{*(2)} > 0$ pour $j = 1, \dots, M^{*(2)}$.

Pour une vague subséquente donnée (que l'on peut assimiler à l'année 2), la population U contient les $M^{(1)}$ individus présents à l'année 1, plus certains individus $M^{(2)}$ absents au départ (c'est-à-dire, absents de la population à l'année 1). La population des individus absents au départ est désignée par $U^{(2)}$. Ainsi, la population $U = U^{(1)} \cup U^{(2)}$ contient $M = M^{(1)} + M^{(2)}$ individus. Si $U^{*(2)}$ représente la population de $M^{*(2)}$ immigrants de l'année 2, nous obtenons $U^{*(2)} \subseteq U^{(2)}$, ainsi que $M^{*(2)} \leq M^{(2)}$. L'astérisque (*) sert ici à rappeler que les nouveaux-nés ont été exclus. Les individus de l'année 2 font partie de N ménages où la taille du ménage i correspond à M_i , $i = 1, \dots, N$.

nous avons choisi pour l'EDTR en janvier 1993 (année 1) un échantillon $s^{(1)}$ de $m^{(1)}$ individus tiré d'une population $U^{(1)}$ de $M^{(1)}$ individus. L'échantillon est choisi dans des logements qui contiennent des ménages. En d'autres mots, les $m^{(1)}$ individus sont obtenus en choisissant $n^{(1)}$ ménages de $N^{(1)}$, chaque ménage i étant assorti d'une probabilité de sélection $\pi_i^{(1)} > 0$, $i = 1, \dots, N^{(1)}$. Désignons par $M^{(1)}$ la taille du ménage i de sorte que $M^{(1)} = \sum_{i=1}^{N^{(1)}} M_i^{(1)}$. Désignons d'autre part par $\pi_j^{(1)}$ la probabilité de sélection de l'individu j . Cette probabilité de sélection est retenue pour l'ensemble des vagues de l'enquête.

Le reste des immigrants sélectionnés aux fins de l'assur-rance d'une représentativité transversale est constitué des individus qui se sont joints aux ménages longitudinaux et qui sont considérés comme des cohabitants. Pour ce qui est des nouveaux-nés et des individus présents au départ dont il est question au paragraphe précédent, l'ajout des cohabitants aux ménages longitudinaux a pour effet d'inclure des individus assortis d'une probabilité de sélection mal définie.

Les individus dont les probabilités de sélection sont mal définies ont été inclus dans le protocole d'enquête d'une manière "illégitime". Ils compliquent la détermination des poids de base à cause du caractère mal défini de leur probabilité de sélection. C'est pour contourner cette difficulté que nous proposons la méthode du partage des poids.

logements, et tous les membres des ménages installés dans ces logements sont interviewés. On utilise un plan de rotation à six groupes pour établir l'échantillon: chaque mois, le groupe qui a fait partie de l'échantillon pendant six mois est remplacé. Chaque groupe de rotation contient environ 10,000 ménages qui comptent environ 20,000 personnes âgées de 16 ans et plus. Pour en savoir plus sur le plan d'échantillonnage de l'EPA, consulter Singh et coll. (1990). Pour l'EDTR, l'échantillon longitudinal ne sera plus actualisé après sa sélection en janvier 1993. Toutefois, pour lui assurer une certaine représentativité transversale, les individus *absents au départ* (c'est-à-dire ceux qui ne faisaient pas partie de la population l'année du choix de l'échantillon longitudinal) devront être pris en compte dans l'échantillon de janvier 1994 et dans les échantillons subséquents. La définition des individus absents au départ comprend les *nouveaux-nés* (nés depuis janvier 1993) et les *immigrants*. À noter que cet ajout à l'échantillon sera transversal du fait que seuls les individus longitudinaux seront inclus de façon permanente dans l'échantillon.

Tableau 1
Terminologie de l'EDTR

Individus: Individus longitudinaux: Individus choisis à l'année 1 dans l'échantillon longitudinal. Individus absents au départ: Individus qui ne faisaient pas partie de la population l'année du choix de l'échantillon longitudinal (année 1). Comprennent les immigrants et les nouveaux-nés. Individus présents au départ: Individus qui faisaient partie de la population à l'année 1, mais qui n'ont pas été choisis à cette époque. Cohabitants: Individus absents ou présents au départ qui se sont joints à un ménage longitudinal. Immigrants: Individus qui, en janvier de l'année 1, se trouvaient à l'extérieur des dix provinces canadiennes et individus vivant dans les zones exclues (Territoires, institutions, réserves indiennes et bases militaires). Nouveaux-nés: Individus nés depuis janvier de l'année 1.	Ménages: Ménages longitudinaux: Ménages qui contiennent au moins un individu longitudinal.
--	--

Nous reproduisons dans le tableau 1 une liste de termes particuliers à l'EDTR. Après la sélection de janvier 1993 (année 1), la population contient les individus longitudinaux et les individus présents au départ. En janvier 1994 (année 2), la population contient les individus longitudinaux, les individus présents au départ et les individus absents au départ. On met l'accent sur les ménages qui contiennent au moins un individu longitudinal (c.-à-d., les *ménages longitudinaux*), les individus absents ou présents au départ qui se joignent à ces ménages étant désignés par le terme *cohabitants*. L'EDTR suivra l'évolution dans le temps des caractéristiques des individus et des ménages. À chaque vague

d'entrevues, tous les membres d'un ménage longitudinal seront questionnés. La composition des ménages longitudinaux changera avec le temps par suite des naissances ou de l'arrivée d'immigrants. La sélection des individus absents au départ pourra être fondée en partie sur les personnes qui se joignent aux ménages longitudinaux.

3.1.2 Échantillon supplémentaire

La restriction imposée concernant les individus absents au départ qui se joignent ensuite à des ménages longitudinaux aura malheureusement pour effet d'exclure les ménages constitués exclusivement d'individus absents au départ (p. ex., les familles d'immigrants). Pour remédier à cette situation, on pourrait entre autres tirer un *échantillon supplémentaire*, par exemple, un échantillon de logements tirés directement de l'EPA en cours à chaque vague d'entrevues. Des questions supplémentaires seraient ainsi ajoutées au questionnaire de l'EPA pour détecter les ménages contenant *au moins un immigrant*, il suffirait ensuite de questionner les membres des ménages ainsi sélectionnés.

En gardant à l'esprit que l'échantillon supplémentaire sert à la sélection de ménages constitués uniquement d'individus absents au départ (immigrants et nouveaux-nés), le fait de limiter cet échantillon aux seuls immigrants ne causera aucun problème de représentativité. On imagine en effet difficilement qu'il puisse exister des ménages uniquement constitués de nouveaux-nés; chaque ménage comprend normalement au moins un adulte. Les nouveaux-nés sont de toute manière déjà représentés dans l'échantillon par les ménages longitudinaux. Par ailleurs, si l'échantillon supplémentaire devait inclure des nouveaux-nés en plus des immigrants, le coût de l'enquête pourrait en être sensiblement augmenté. En effet, l'échantillon supplémentaire inclurait un ménage complet pour chaque nouveau-né sélectionné, grossissant ainsi inutilement l'échantillon et les coûts de l'échantillonnage puisque les nouveaux-nés sont déjà représentés dans l'échantillon.

Au lieu d'avoir recours à l'EPA en cours, on pourrait également envisager de tirer l'échantillon supplémentaire en visitant à nouveau les logements utilisés lors de la sélection de l'échantillon initial. Cette méthode présente certains avantages pratiques (par exemple, il est plus facile de se rendre à des adresses déjà connues). Elle poserait cependant le problème des nouveaux logements qui n'existaient pas en janvier 1993. La probabilité de sélection de ces nouveaux logements dans l'échantillon supplémentaire serait nulle, ce qui constituerait une source de biais. C'est là une des raisons pour lesquelles la première méthode nous paraît préférable (détection des ménages contenant au moins un immigrant au moyen du questionnaire de l'EPA en cours).

La figure 2 illustre brièvement la sélection longitudinale et transversale des individus. Les lettres et les maisons y représentent respectivement les individus et les ménages. Les individus A, D, E et F sont des individus longitudinaux qui sont suivis dans le temps. L'individu C est un individu présent au départ, c'est-à-dire qui faisait partie de la population à l'année 1 mais qui n'a pas été choisi à ce moment.

$$Y = \sum_{M^A}^{M^A} \frac{t_j}{t_j} \sum_{M^B}^{M^B} \pi_{jj'}^A \sum_{k=1}^{k=1} l_{jk} z_k \quad (5)$$

En considérant maintenant l'espérance, nous obtenons

$$E(Y) = \sum_{M^A}^{M^A} \frac{E(t_j)}{\pi_{jj'}^A} Z_j$$

puisque $E(t_j) = \pi_{jj'}^A$.

Il suffit maintenant de démontrer que $Z = Y$. Nous avons

d'abord

$$Z = \sum_{M^A}^{M^A} Z_j = \sum_{M^A}^{M^A} \sum_{k=1}^{k=1} l_{jk} z_k = \sum_{M^B}^{M^B} z_k \sum_{j=1}^{j=1} l_{jk}$$

En reformulant cette somme en fonction des N grappes de la population U_B , nous obtenons

$$Z = \sum_{M^B}^{M^B} \sum_{k=1}^{k=1} z_{ik} \sum_{M^A}^{M^A} l_{jik} = \sum_{M^B}^{M^B} z_{ik} \sum_{j=1}^{j=1} l_{jik}$$

$$= \sum_{M^B}^{M^B} \sum_{k=1}^{k=1} \frac{L_i}{Y_i} L_{ik} = \sum_{N}^{N} Y_i = Y.$$

Le caractère non biaisé de la méthode du partage des poids peut également être prouvé à l'aide d'une méthode semblable à celle présentée par Ernst (1989).

2.2 Estimation de la variance

Pour obtenir une formule de la variance pour Y , nous partons de l'équation (5). Puisqu'il s'avère que Y n'est rien d'autre qu'un estimateur de Z de Horvitz-Thompson (voir Horvitz et Thompson 1952), la variance de Y est donnée directement par

$$\text{Var}(Y) = \sum_{M^A}^{M^A} \sum_{j'=1}^{j'=1} \frac{(\pi_{jj'}^A - \pi_{jj'}^A \pi_{jj'}^A)}{Z_j Z_{j'}} \quad (6)$$

où $\pi_{jj'}^A$ représente la probabilité conjointe de sélection des unités j et j' (voir Särndal, Swensson et Wretman (1992) pour le calcul de $\pi_{jj'}^A$ sous divers plans d'échantillonnage).

3.1 Plan d'échantillonnage

3.1.1 Échantillon initial

L'échantillon longitudinal de l'EDTR a été tiré en janvier 1993 à partir de deux groupes supprimés par renouvellement de l'Enquête sur la population active (EPA); il s'agissait donc d'un sous-échantillon de l'EPA. Il est constitué de près de 15,000 ménages. Le terme ménage désigne ici toute personne ou groupe de personnes vivant dans un logement; il peut s'agir d'une personne vivant seule, d'un groupe de personnes non apparentées qui partagent le même logement ou des membres d'une famille. L'EPA est une enquête continue conçue pour fournir des estimations mensuelles du nombre de travailleurs salariés, de travailleurs autonomes et de chômeurs. Elle utilise un plan d'expérience stratifié à plusieurs degrés fondé sur une base de sondage aréolaire où les logements constituent les unités finales d'échantillonnage. L'échantillon de l'EPA est constitué de toutes les personnes membres des ménages qui occupent les logements sélectionnés. En d'autres mots, l'EPA tire un échantillon de

3. APPLICATION À L'EDTR

L'EDTR a été lancée par Statistique Canada en janvier 1994. Cette enquête a pour objectif d'observer l'évolution dans le temps de l'activité des individus dans le marché du travail, ainsi que les changements qui surviennent dans le revenu individuel et la situation des ménages. L'EDTR vise avant tout à fournir des données longitudinales, mais elle produira également des estimations transversales. La population cible de l'EDTR est l'ensemble des individus qui vivent dans les provinces du Canada, sans distinction d'âge. Pour des raisons opérationnelles, les habitants des Territoires et les individus qui vivent en institutions, dans les réserves indiennes et dans les bases militaires sont exclus de l'échantillon (voir Lavalée 1993).

$$\widehat{\text{Var}}(Y) = \sum_{M^A}^{M^A} \sum_{j'=1}^{j'=1} \frac{(\pi_{jj'}^A - \pi_{jj'}^A \pi_{jj'}^A)}{Z_j Z_{j'}} \quad (7)$$

de l'équation suivante:

La variance $\text{Var}(Y)$ peut être estimée sans biais à l'aide de l'estimateur de Horvitz-Thompson. L'estimateur de Horvitz-Thompson substitue chaque Z_j dans l'équation de la variance de puis de calculer $Z_j = \sum_{k=1}^{M^B} l_{jk} z_k$. Il ne reste plus qu'à d'abord de calculer $z_k = Y_i / L_i$ pour chaque unité $k \in i$, En pratique, l'équation (6) est facile à utiliser. Il suffit

La méthode d'estimation que nous présentons utilise l'échantillon s^A ainsi que les liens qui existent entre U^A et U^B pour donner une estimation du total Y appartenant à la population U^B . Les liens servent en fait de ponts qui nous permettent d'aller de la population U^A à la population U^B , et vice versa. À signaler qu'en pratique, il pourrait s'avérer physiquement impossible de tirer directement un échantillon s^B de U^B comme nous l'avons décrits dans l'exemple présenté à l'introduction.

Pour estimer le total Y , on peut utiliser l'estimateur

Pour estimer le total Y , on peut utiliser l'estimateur

$$(I) \quad \sum_{k=1}^M \sum_{l=1}^n w_{lk} y_{lk} = \chi$$

ou n représente le nombre de grappes interviewées, et w_{ik} le poids rattaché à l'unité k de la grappe i . Pour obtenir des estimations non biaisées, on peut utiliser des poids correspondant à l'inverse des probabilités de sélection des unités qui entrent dans l'estimateur \bar{Y} . Pour chaque unité k de la grappe i ayant un lien $l_{j,k} = 1$ avec une unité j de U_A , ce choix est possible puisque $\pi_k^B = \pi_j^A$. Toutefois, les unités de U_B n'ont pas nécessairement toutes un lien avec U_A . En outre, même si un lien existait, rien ne garantit

En règle générale, la méthode du partage des poids attribuée à chaque unité échantillonnée un poids de base établi à partir d'une moyenne calculée au sein de chaque grappe i entrant dans Y . On obtient d'abord un *poids initial* correspondant à l'inverse de la probabilité de sélection pour l'unité k de la grappe i de Y ayant un lien $f_{i,k} = 1$ avec une unité $j \in s^A$. Un poids initial de zéro est attribué aux unités qui n'ont pas de lien. Le *poids de base* s'obtient en calculant la moyenne des poids initiaux de la grappe. À signaler que le fait d'attribuer le même poids de base à toutes les unités présente l'énorme avantage d'assurer la cohérence des estimations au niveau des unités et des grappes.

Selon cette méthode, chaque unité k d'une grappe i entrant dans Y se voit attribuer un poids initial $w_{i,k}$ comme suit :

Selon cette méthode, chaque unité k d'une grappe i entrant dans Y se voit attribuer un poids initial w_{ik} comme suit:

$$w_{ik}' = \sum_{j=1}^{M^A} l_{f,jk} \frac{y_{fj}}{l_{fj}}, \quad (2)$$

où $t_j = 1$ si $j \in s^A$, et 0 dans les autres cas. À signaler qu'une unité k n'ayant de lien avec aucune des unités j de U^A se voit attribuer automatiquement un poids initial de zéro.

et finalement,

$$\hat{Y} = \sum_{k=1}^K \sum_{j=1}^{M_A} l_{jk} \frac{\pi_j}{t_j} z_k$$

Cependant, puisque $t_j \neq 0$ seulement pour les unités k qui entrent dans Y , nous pouvons étendre la première somme à toutes les unités k dans U^B . Ainsi,

$$\sum_{k=1}^K w_k' z_k = \sum_{k=1}^K \left[\sum_{j=1}^M l_{jk} \frac{\pi_j}{t_j} \right] z_k.$$

Utilisons maintenant un indice unique k pour désigner les unités m^B qui entrent dans $Y(m^B = \sum_{i=1}^n M_i^B)$. En remplaçant w_i^k par sa définition (2), nous obtenons

$$Y = \sum_{n=1}^{\infty} \sum_{k=1}^{M_n} w'_{ik} z_{ik}. \quad (4)$$

Si $z^{i,k} = Y_i/L_i$ pour tous les $k \in i$, nous obtenons alors

$$Y = \sum_{i=1}^n Y_i = \left[\frac{\sum_{k=1}^{M_B^i} L_{i^k}}{\sum_{k=1}^{M_B^i} w_{i^k}} \right] = \sum_{i=1}^n \frac{Y_i}{L_i} \sum_{k=1}^{M_B^i} L_{i^k} w_{i^k}.$$

Nous allons maintenant démontrer que l'estimateur \hat{Y} est non biaisé pour Y avec la méthode du partage des poids. À partir de $\hat{Y} = \sum_{i=1}^n w_i^t \sum_{k=1}^K M_{ik}^B$, nous substituons la définition de w_i^t dans \hat{Y} pour obtenir

2.1 Caractère non biaisé de la méthode du partage des poids

sein de la grappe i .
Finalement, nous posons que $w_{ik} = w_i$ pour tous les $k \in I_i$.

$$(3) \quad \frac{\sum_{k=1}^{M_B^I} T_{ik}}{\sum_{k=1}^{M_B^I} w_{ik}} = w_i$$

l'EDTR. Il s'agit d'un exemple des enquêtes longitudinales pour lesquelles la production d'estimations transversales à partir d'un échantillon longitudinal présente d'importantes difficultés. L'EDTR est une enquête longitudinale typique menée auprès des individus et des ménages. Troisièmement, nous décrirons l'utilisation d'un échantillon supplémentaire pour améliorer la représentativité transversale de l'échantillon longitudinal initial. Quatrièmement, nous aborderons la notion des poids de base, équivalents ici aux poids de sondage. Finalement, nous décrirons l'utilité de la méthode du partage des poids pour le calcul des poids de base de tous les individus interviewés dans le cadre de l'EDTR.

2. LA MÉTHODE DU PARTAGE DES POIDS – APERÇU GÉNÉRAL

Ernst (1989) décrit la méthode du partage des poids dans le contexte des enquêtes longitudinales menées auprès des ménages. Dans le même contexte, Kalton et Brick

(1995) examinent différents plans de pondération, et notamment la méthode du partage des poids. Gailly et Lavalée (1993) décrivent pour leur part diverses conséquences de l'utilisation de cette méthode dans le cadre d'enquêtes longitudinales menées auprès des ménages. Nous examinerons cette méthode sous l'angle général des situations d'échantillonnage où la population visée doit être échantillonnée en vertu d'une base de sondage se rapportant à une population différente, mais liée d'une certaine façon à la première. À signaler qu'on peut assimiler cette situation à celle d'un sondage de réseaux (voir Thompson 1992). Par exemple, on peut imaginer une situation où l'échantillon doit porter sur les jeunes enfants alors qu'on ne dispose pour toute base de sondage que d'une liste de noms de parents. La population visée est celle des enfants, mais nous devons d'abord tirer un échantillon de parents avant de pouvoir sélectionner l'échantillon d'enfants. À signaler que les enfants d'une famille particulière peuvent être échantillonnés par l'intermédiaire du père ou de la mère. La situation où on souhaite mener une enquête auprès des entreprises alors qu'on ne dispose que d'une base de sondage incomplète des établissements est un autre exemple d'application de cette méthode. Pour chaque établissement tiré de la base de sondage, on voudra échantillonner l'ensemble des établissements appartenant à la même entreprise. Les établissements qui ne figurent pas dans la base devront être représentés par ceux qui font partie de cette base.

Supposons qu'un échantillon s^A de m^A unités est tiré d'une population U^A de M^A unités dans le cadre d'un plan d'échantillonnage quelconque. Désignons par π_j^A la probabilité de sélection de l'unité j . Nous présumons que $\pi_j^A > 0$ pour tous les $j \in U^A$. Désignons par U^B une population de M^B unités. Cette population est divisée en N grappes, dont la grappe i contient M_i^B unités. Par exemple, dans le contexte des enquêtes sociales, les grappes peuvent représenter des ménages, et les unités peuvent représenter les individus qui font partie de ces ménages. De même, pour les enquêtes menées auprès des entreprises, les grappes peuvent représenter des entreprises,

mesures de lutte à la pauvreté. Nous présumons qu'il existe un lien (ou une correspondance) entre chaque unité j de la population U^A et au moins une unité k de la population U^B . En outre, chaque grappe i de U^B de possède au moins un lien avec une unité j de U^A . Ce lien est déterminé par une variable indicatrice I_{jk} , où $I_{jk} = 1$ s'il existe un lien entre l'unité $j \in U^A$ et l'unité $k \in U^B$, et 0 dans les autres cas. Toutes les unités de la population U^A ont au moins un lien avec la population U^B , ainsi, $L_j^A = \sum_{k \in U^B} I_{jk} \geq 1$ pour tous les $j \in U^A$. Toutefois, le nombre de liens pour une unité k de la population U^B peut être égal à zéro, à un ou à un nombre supérieur à un. Ainsi, on peut avoir $L_k^B = \sum_{j \in U^A} I_{jk} = 0$ ou $L_k^B = \sum_{j \in U^A} I_{jk} > 1$ pour certains $k \in U^B$. Cette situation est illustrée à la figure 1.

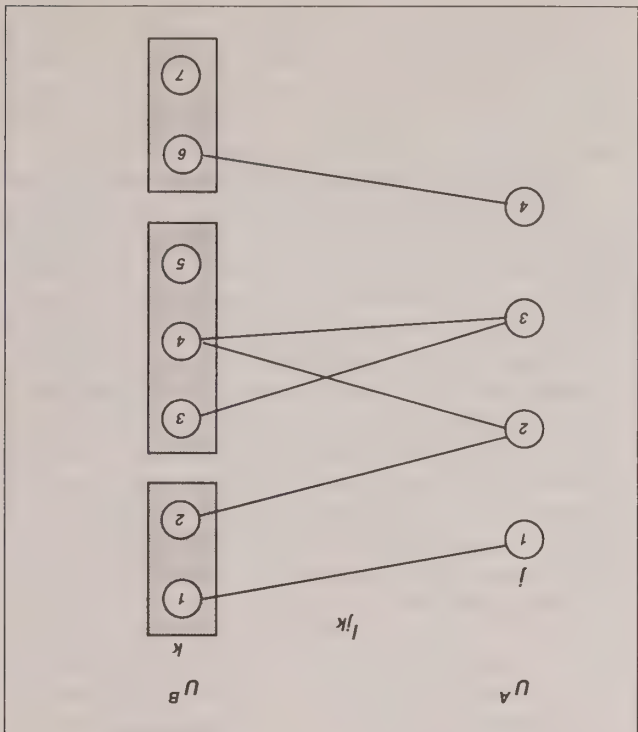


Figure 1. Liens entre les unités des populations U^A et U^B .

Pondération transversale des enquêtes longitudinales menées auprès des individus et des ménages à l'aide de la méthode du partage des poids

PIERRE LAVALLÉE¹

RÉSUMÉ

Les organismes de statistiques effectuent de plus en plus d'enquêtes longitudinales. Or, même si ces enquêtes produisent surtout des données longitudinales, on s'attend dans la plupart des cas qu'elles donnent également des estimations transversales fiables. Les enquêtes menées auprès des individus et des ménages doivent tenir compte de la dynamique des populations, c'est-à-dire des changements sensibles qui peuvent survenir, avec le temps, dans la composition des ménages. Ainsi, il convient d'adapter les méthodes d'estimation transversale à la nature longitudinale de l'échantillon. Dans le présent document, nous présentons un aperçu général de la méthode du partage des poids, laquelle peut servir entre autres à attribuer un poids de base aux membres d'un ménage. Nous nous penchons également sur l'estimateur de variance associé à cette méthode. Nous examinons finalement le problème que pose la pondération d'un échantillon longitudinal dans les cas où un échantillon supplémentaire est sélectionné pour améliorer la représentativité transversale de l'échantillon. Nous examinons, en guise d'application, l'Enquête sur la dynamique du travail et du revenu (EDTR) mise en oeuvre par Statistique Canada en 1994. Cette enquête longitudinale s'intéresse à l'expérience pratique des individus, à l'évolution de leurs revenus et aux changements dans la composition des ménages.

MOTS CLÉS: Méthode du partage des poids; enquête longitudinale; estimation transversale; échantillon supplémentaire.

1. INTRODUCTION

Les organismes de statistiques s'intéressent de plus en plus aux enquêtes longitudinales, c'est-à-dire celles qui portent sur l'évolution des unités d'enquête avec le temps. Statistique Canada travaille actuellement à l'élaboration de trois de ces enquêtes destinées aux individus: l'Enquête nationale sur la santé de la population, l'Enquête longitudinale nationale sur les enfants et l'Enquête sur la dynamique du travail et du revenu (EDTR). Ces enquêtes ont pour objectif principal d'obtenir des données longitudinales qui pourront servir, entre autres, à étudier les changements subis par les variables avec le temps (p. ex., les données longitudinales peuvent servir à analyser la nature chronique de la pauvreté). Elles peuvent en second lieu fournir des estimations transversales, c'est-à-dire des estimations représentatives de l'état d'une population à un point donné dans le temps. Même si elles sont beaucoup moins importantes que les données longitudinales, ces estimations constituent, pour beaucoup d'utilisateurs, un aspect essentiel des enquêtes. Un aperçu transversal représentatif de la population actuelle peut en effet s'avérer utile pour mesurer l'évolution de l'état de la population avec le temps. La nature longitudinale de l'enquête améliore par ailleurs l'exactitude de la mesure du changement.

Nous nous proposons d'étendre l'application de la méthode du partage des poids présentée dans Ernst (1989). Même si cette méthode a vu le jour dans le contexte particulier des enquêtes longitudinales menées auprès des ménages, nous montrerons qu'il est possible de l'appliquer

à des situations où la base de sondage se rapporte à une population différente, quoique liée d'une certaine façon à la population qui nous intéresse. Dans le contexte des enquêtes longitudinales, la base de sondage peut se rapporter à la population initiale alors que la population visée est celle qui existera quelques années plus tard. Nous présentons en outre une nouvelle preuve du caractère non biaisé de la méthode du partage des poids, ainsi que la formule et l'estimateur de la variance qu'il convient d'utiliser avec cette méthode.

Lorsqu'on utilise la méthode du partage des poids, il convient de s'assurer que l'échantillon longitudinal peut servir à une estimation transversale. La difficulté réside dans le fait que même si l'échantillon longitudinal demeure constant, la distribution de la population (individus et ménages) change avec le temps. Au niveau des individus, ces changements sont le fait d'événements tels que les naissances et les décès, l'immigration et l'émigration, et les déménagements d'une région à l'autre du pays. La naissance ou le décès d'une personne modifie bien évidemment la composition d'un ménage, et des événements comme le mariage, le divorce, la séparation, le départ d'un enfant de la population au sein du ménage. Pour obtenir une estimation transversale exacte et non biaisée à partir d'un échantillon longitudinal, il convient donc d'utiliser une méthode qui tiendra compte de tous ces changements.

Nous nous attachons tout d'abord à examiner la méthode du partage des poids sous un angle général. Deuxièmement, nous présenterons le plan d'échantillonnage de

¹ Pierre Lavallée, Division des méthodes d'enquêtes sociales, Statistique Canada, Ottawa (Ontario), Canada, KIA 0T6.

Il va de soi que la méthode III est plus efficace que la méthode I.

Nous savons que $E(1/d) \geq 1/n$, autrement dit $V_{III}(\hat{\pi}) > V_{II}(\hat{\pi})$, ce qui implique que la méthode III est plus efficace que la méthode II.

Comme $1/E(d) \geq 1/n$, la méthode III est plus efficace que la méthode IV.

Il reste à comparer les méthodes I et IV. Comme $E(1/d) > 1/E(d)$ pour $n > 1$, nous obtenons à partir de l'équation (8)

$$\frac{N/E(d) - 1}{1} \leq \frac{N - 1}{n}$$

ce qui implique que en (7) et en (2),

$$\frac{N - E(d)}{\pi(1 - \pi)} \frac{1 - E(d)}{\pi} \leq \frac{N - 1}{\pi(1 - \pi)}$$

pour $n > 1$. En outre, $1/E(d) \geq 1/n$. Cela montre que le second terme du membre de droite de l'équation (7) aura une valeur plus élevée que le terme correspondant de l'équation (2). Par conséquent, les efficacités relatives des méthodes I et IV dépendent de la valeur relative de π et de p . À titre d'exemple, si $N = 100$, $n = 10$ et $p = 0.9$, la méthode IV sera plus efficace que la méthode I pour $0.18 \leq \pi \leq 0.82$.

BIBLIOGRAPHIE

- FRANKLIN, L.A. (1989). Echantillonnage pour populations dichotomiques par la méthode des réponses randomisées avec randomisation continue. *Techniques d'enquête*, 15, 235-245.
- SETH, G.R., et RAO, J.N.K. (1964). On the comparison between simple random sampling with and without replacement. *Sankhyā (A)*, 26, 85-86.
- WARNER, S.L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60, 63-69.
- KIM, J.-I., et FLUECK, J.A. (1978). Modifications of the randomized response technique for sampling without replacement. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 346-350.
- KORWAR, R.M., et SERFLING, R.J. (1970). On averaging over distinct units in sampling with replacement. *Annals of Mathematical Statistics*, 41, 2132-2134.
- KUK, A.Y.C. (1990). Asking sensitive questions indirectly. *Biometrika*, 77, 436-438.
- MANGAT, N.S. (1994). An improved randomized response strategy. *Journal of the Royal Statistical Society, Series B*, 56, 93-95.
- MANGAT, N.S., et SINGH, R. (1990). An alternative randomized response procedure. *Biometrika*, 77, 439-442.
- MANGAT, N.S., et SINGH, R. (1991). An alternative approach to randomized response survey. *Statistica*, anno LI, 327-332.
- MANGAT, N.S., SINGH, R., et SINGH, S. (1992). An improved unrelated question randomized response strategy. *Calcutta Statistical Association Bulletin*, 42, 277-281.
- MURTHY, M.N. (1967). *Sampling Theory and Methods*. Calcutta, India: Statistical Publishing Society.

2. COMPARAISONS D'EFFICACITÉ

Korwar et Serfling (1970) ont montré que, pour $n \geq 3$,

$$\bar{Q} - \frac{1}{1} \frac{720N}{1} < E\left(\frac{d}{1}\right) \leq \bar{Q}$$

où

$$\bar{Q} = \frac{1}{1} + \frac{1}{2N} + \frac{n}{n-1} \frac{1}{12N^2}.$$

Examinons maintenant la formule de la variance en (5). En utilisant \bar{Q} , on vérifie facilement que

$$NE_1(1/d) - 1 \leq \frac{N-1}{1}, \quad (8)$$

dans le premier terme du membre de droite de l'équation (5) mais que $E_1(1/d) \geq 1/n$ dans le second terme. Par conséquent, l'efficacité relative de l'estimateur défini en (1) - EASAR avec unités répétées - par rapport à l'estimateur défini en (3) - EASAR avec unités non répétées - dépendra de la valeur relative de π et de p . Cela s'explique par le fait que les unités répétées peuvent produire des réponses différentes à cause du mécanisme de randomisation et, par conséquent, générer de l'information additionnelle. On définit une condition suffisante pour que l'inéquation $V_{II}(\hat{\pi}_d) - V_I(\hat{\pi}) < 0$ soit vérifiée en utilisant $E_1(d) = \bar{Q}$. Ainsi, on obtient la condition suivante:

$$\pi(1-\pi) > \frac{n(N-1)(6N+n-1)}{d(1-p)} \frac{N\{6Nn-12N-n(n-1)\}}{(2p-1)^2}. \quad (9)$$

L'inéquation ci-dessus se vérifiera probablement pour des valeurs de p qui sont proches de 0 ou de 1, c.-à-d. les cas où il y a un risque d'associer le répondant à un groupe particulier. Par exemple, si $N = 100$, $n = 10$ et $p = 0.9$, l'inéquation (9) se vérifiera pour $0.236 \leq \pi \leq 0.764$. De même, la méthode II sera inférieure à la méthode I si $V_{II}(\hat{\pi}_d) - V_I(\hat{\pi}) > 0$. En utilisant $E_1(1/d) = \bar{Q} - 1/720N$, on obtient l'inéquation

$$\pi(1-\pi)$$

$$< \frac{n(N-1)\{359N+60(n-1)\}}{d(1-p)} \frac{N\{361Nn-720N-60n(n-1)\}}{(2p-1)^2}.$$

Si on prend le même exemple que précédemment, cette inéquation se vérifiera pour $\pi \leq 0.234$ ou $\pi \geq 0.764$. Si nous utilisons l'inégalité de Cauchy-Schwarz, $E(1/d) > 1/E(d)$, comme dans Seth et Rao (1964), nous constatons que $V_{II}(\hat{\pi}_d) > V_{IV}(\hat{\pi}_d)$, ce qui signifie que la méthode IV est plus efficace que la méthode II.

1.3 Méthode III

On tire un échantillon de n répondants suivant un plan EASSR (Kim et Flueck 1978). Dans ce cas, on reprend l'estimateur $\hat{\pi}$ défini en (1) et on peut exprimer la variance correspondante en remplaçant d par n dans l'équation (4), ce qui donne

$$V_{III}(\hat{\pi}) = \frac{N-n}{N-1} \frac{\pi(1-\pi)}{d(1-p)} + \frac{n(2p-1)^2}{d(1-p)}. \quad (6)$$

1.4 Méthode IV

Dans ce cas-ci, l'estimateur repose sur un EASSR de taille $E(d)$. Le coût prévu est le même pour l'EASSR et l'EASSR. Pour ce plan, l'estimateur sera

$$\hat{\pi}_E = \frac{d'/E(d) - 1 + p}{2p - 1}, \quad p \neq .5$$

et la variance correspondante,

$$V_{IV}(\hat{\pi}_E) = \frac{N/E(d) - 1}{N - 1} \frac{\pi(1-\pi)}{d(1-p)} + \frac{E(d)(2p-1)^2}{d(1-p)^2}. \quad (7)$$

Si E_1 et V_1 désignent respectivement l'espérance et la variance pour l'ensemble des valeurs de d , nous avons alors $V_{II}(\hat{\pi}_d) = E_1 V_2(\hat{\pi}_d) + V_1 E_2(\hat{\pi}_d)$. En utilisant l'équation (4), on obtient

$$V_{II}(\hat{\pi}_d) = \left[NE_1\left(\frac{d}{1}\right) - 1 \right] \frac{\pi(1-\pi)}{d(1-p)} + E_1\left(\frac{d}{1}\right) \frac{(2p-1)^2}{d(1-p)^2}. \quad (5)$$

puisque le second terme du membre de droite de l'équation de $V_{II}(\hat{\pi}_d)$ est égal à zéro si $E_2(\hat{\pi}_d) = \pi$.

Etant donné d unités distinctes, l'échantillon équivalait à un échantillon aléatoire simple sans remise de taille d tiré parmi N unités. L'estimateur $\hat{\pi}_d$ est donc non biaisé pour la proportion de population π . Pour analyser l'efficacité de l'estimateur proposé $\hat{\pi}_d$, nous avons besoin de la variance. Voici l'expression pour la variance conditionnelle $V_2(\hat{\pi}_d)$ pour une valeur donnée de d . Ainsi,

$$V_2(\hat{\pi}_d) = \frac{N-d}{N-1} \frac{\pi(1-\pi)}{d(1-p)} + \frac{d(2p-1)^2}{d(1-p)^2}. \quad (4)$$

$$\hat{\pi}_d = \frac{d'/d - 1 + p}{2p - 1}, \quad p \neq .5. \quad (3)$$

Efficacité de l'emploi des unités non répétées dans une enquête fondée sur la méthode des réponses randomisées

N.S. MANGAT, R. SINGH, S. SINGH, D.R. BELLHOUSE et H.B. KASHANI¹

RÉSUMÉ

Tout le monde sait que, dans un échantillonnage aléatoire simple avec remise, la moyenne d'échantillon calculée uniquement sur la base des unités non répétées est plus efficace que la moyenne d'échantillon calculée sur la base de toutes les unités échantillonnées sans distinction (Murthy 1967, pp. 65-66). Seth et Rao (1964) ont montré que, étant donné le même coût moyen d'échantillonnage, la moyenne d'échantillon calculée dans un échantillonnage répétées dans un échantillonnage avec remise était moins efficace que la moyenne calculée dans un échantillonnage sans remise. En nous servant de la méthode des réponses randomisées de Warner (1965), nous comparons l'échantillonnage aléatoire simple sans remise et l'échantillonnage avec remise où seules les unités non répétées sont prises en compte.

MOTS CLÉS: Échantillonnage aléatoire simple avec remise et sans remise; inférence avec unités non répétées; méthode de Warner.

1. INTRODUCTION

C'est à Warner (1965) que l'on doit l'introduction de la méthode des réponses randomisées (RR); cette méthode permet d'obtenir des données fiables lorsqu'on cherche à estimer la proportion π de la population possédant une caractéristique délicate à révéler. Depuis la parution de l'article de Warner, beaucoup de chemin a été parcouru. Mentionnons notamment, parmi les plus récents, les ouvrages de Franklin (1989), Kuk (1990), Mangat et Singh (1990, 1991), Mangat, Singh et Singh (1992) et Mangat (1994), qui proposent tous d'autres méthodes ou estimateurs fondés sur la méthode des RR.

On sait que, dans les enquêtes à plan d'échantillonnage aléatoire simple avec remise (EASAR), l'estimateur de la moyenne de population basé sur les unités non répétées est toujours plus efficace que l'estimateur de la moyenne basé sur toutes les unités échantillonnées (Murthy 1967, pp. 65-66). En outre, Seth et Rao (1964) ont montré que, étant donné le même coût moyen d'échantillonnage, la moyenne d'échantillon calculée dans un échantillonnage sans remise est plus efficace que la moyenne calculée sur la base des unités non répétées dans un échantillonnage avec remise. Nous avons donc voulu savoir si les observations précédentes s'appliquent aussi au modèle original des réponses randomisées de Warner, qui sert couramment, dans la pratique, à l'échantillonnage de répondants pour des enquêtes qui portent sur des sujets délicats. Pour analyser le problème, nous allons considérer l'utilisation de quatre méthodes d'échantillonnage.

Selon cette méthode, qui est en fait la méthode de Warner, on présente à chaque répondant d'un échantillon

1.1 Méthode I

On tire un échantillon de n répondants dans un population finie de N unités suivant un plan EASAR, sauf que l'information tirée des d unités non répétées de l'échantillon, $1 \leq d \leq n$, est utilisée dans la construction de l'estimateur. Posons d' comme le nombre d'enquêtes qui ont répondu "oui" dans l'interview faite au moyen d'un dispositif de randomisation. Nous considérons alors l'estimateur suivant pour π :

1.2 Méthode II

On devrait choisir une valeur de p aussi près que possible de 1 ou de 0 sans pour autant compromettre la collaboration des répondants.

$$V_1(\hat{\pi}) = \frac{\pi(1-\pi)}{\pi(1-p)} + \frac{n}{p(1-p)} \cdot \frac{n(2p-1)^2}{2} \quad (2)$$

est non biaisé pour π et sa variance est définie par l'expression

$$\hat{\pi} = \frac{n'/n - 1 + p}{2p - 1}, \quad p \neq .5, \quad (1)$$

aléatoire simple avec remise (EASAR) un dispositif de randomisation qui consiste en deux énoncés du type: i) "J'appartiens au groupe auquel l'enquête s'intéresse" et ii) "Je n'appartiens pas au groupe auquel l'enquête s'intéresse", la probabilité p étant rattachée au premier énoncé et la probabilité $(1-p)$, au second. L'enquête répond "oui" ou "non" à l'énoncé tiré au hasard suivant sa situation, sans révéler l'énoncé. Si n' personnes de l'échantillon (y compris les unités répétées) ont répondu "oui", l'estimateur de Warner

variables aléatoires. Comme dans le cas du modèle des constantes additives, on présume que le répondant utilise u' et v' au lieu de u et v . Ensuite, en prenant séparément les espérances du numérateur et de l'expression placée sous chacun des signes de racine carrée dans le dénominateur de (19), on obtient l'expression

$$\sigma_{xy} = \frac{\sqrt{\sum_{j=1}^N x_j^2 \frac{\sigma_u^2 \mu_{u'}^2 - \sigma_v^2 \mu_{v'}^2}{\sigma_u^2 + \mu_{u'}^2}} + \sqrt{\sum_{j=1}^N y_j^2 \frac{\sigma_v^2 \mu_{v'}^2 - \sigma_u^2 \mu_{u'}^2}{\sigma_v^2 + \mu_{v'}^2}}}{\sqrt{\sum_{j=1}^N x_j^2 \frac{\sigma_u^2 \mu_{u'}^2}{\sigma_u^2 + \mu_{u'}^2}} + \sqrt{\sum_{j=1}^N y_j^2 \frac{\sigma_v^2 \mu_{v'}^2}{\sigma_v^2 + \mu_{v'}^2}}} \quad (24)$$

Si $\mu_u = \mu_{u'}$, $\mu_v = \mu_{v'}$, $\sigma_u^2 \geq \sigma_v^2$ et $\sigma_{u'}^2 \geq \sigma_{v'}^2$, comme c'est le cas avec le modèle des constantes additives, il en découle que selon l'expression (24), le biais dans les réponses conduit à une surestimation de la corrélation.

Dans le cas du modèle des questions non liées, on peut raisonnablement présumer, quand on se penche sur la question du biais dans les réponses, que les répondants répondent aux questions délicates avec une probabilité $p_1' < p_1$ et $p_2' < p_2$. En général, l'effet de ce biais dépend des valeurs relatives des diverses probabilités, des moyennes et des variances des questions délicates, et des moyennes et des variances des questions non délicates. En vertu du plan d'échantillonnage aléatoire simple sans remise et du modèle du biais dans les réponses, la valeur de l'espérance du numérateur de (20) est donnée par

$$p_1' p_2' \left[S_{xy} + \frac{p_1' p_2'}{(1 - p_1') (1 - p_2') - (1 - p_1) (1 - p_2)} S_{uv} \right]$$

qui est plus grand que $p_1' p_2' S_{xy}$. De la même façon, l'espérance de l'expression S_{xy}^2 dans (20) est donnée par

$$S_{xy}^2 [p_1'^2 + (N - 1) p_1' (p_1 - p_1') / N] + (p_1 - p_1') S_{xy}^2 [p_1' - (p_1' + 2p_1 - 2) / N] + p_1' (p_1 - p_1') (X - U)^2,$$

qui est plus grand que $p_1'^2 S_{xy}^2$ lorsque N est grand. Si $S_{uv} = 0$, le biais dans les réponses a tendance à sous-estimer la corrélation.

REMERCIEMENTS

Cet article est dédié à la mémoire de Stan Warner.

Il a été préparé en vue de la séance du *Stanley Warner Memorial* tenue à l'occasion des rencontres de la Société statistique du Canada tenues à Banff (Alberta), en mai 1994. Ces travaux ont bénéficié d'une aide financière du Conseil de recherches en sciences naturelles et en génie du Canada.

BIBLIOGRAPHIE

- BELLHOUSE, D.R. (1980). Linear models for randomized response designs. *Journal of the American Statistical Association*, 75, 1001-1004.
- CHAUDHURI, A., et MUKERJEE, R. (1988). *Randomized Response: Theory and Techniques*. New York: Marcel Dekker.
- EDGELL, S.E., HIMMELFARB, S., et CIRIA, D.J. (1986). Statistical efficiency of using two quantitative randomized response techniques to estimate correlation. *Psychological Bulletin*, 100, 251-256.
- GREENBERG, B.G., ABUL-ELA, A.A., SIMMONS, W.R., et HORVITZ, D.G. (1969). The unrelated question randomized response model: theoretical framework. *Journal of the American Statistical Association*, 64, 520-539.
- GREENBERG, B.G., KUEBLER, R.R., ABERNATHY, J.R., et HORVITZ, D.G. (1971). Application of the randomized response technique in obtaining quantitative data. *Journal of the American Statistical Association*, 66, 243-250.
- LEYSIEFFER, F.W., et WARNER, S.L. (1976). Respondent jeopardy and optimal designs in randomized response models. *Journal of the American Statistical Association*, 71, 649-656.
- NATHAN, G. (1988). Bibliographie de la méthode des réponses randomisées, 1965-1987. *Techniques d'enquête*, 14, 351-365.
- POLLOCK, K.H., et BEK, Y. (1976). A comparison of three randomized response models for quantitative data. *Journal of the American Statistical Association*, 71, 884-886.
- RAO, C.R. (1952). Some theorems on minimum variance unbiased estimation. *Sankhyā* (A), 12, 27-42.
- RAO, J.N.K. (1975). On the foundations of survey sampling. Dans *A Survey of Statistical Design and Linear Models*. (Ed. J.N. Srivastava). Amsterdam: North-Holland, 489-505.
- RAO, J.N.K., et BELLHOUSE, D.R. (1978). Optimal estimation of a finite population mean under generalized random permutation models. *Journal of Statistical Planning and Inference*, 2, 125-141.
- STEM, D.E., et STEINHORST, R.K. (1984). Telephone interview and mail questionnaire applications of the randomized response model. *Journal of the American Statistical Association*, 79, 555-564.
- UMESH, U.N., et PETERSON, R.A. (1991). A critical evaluation of the randomized response method. *Sociological Methods and Research*, 20, 104-138.
- WARNER, S.L. (1965). Randomized response: a survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60, 63-69.
- WARNER, S.L. (1971). The linear randomized response model. *Journal of the American Statistical Association*, 66, 884-888.
- WARNER, S.L. (1976). Optimal randomized response models. *Revue Internationale de Statistique*, 44, 205-212.
- WARNER, S.L. (1986). The omitted digit randomized response model for telephone applications. *Proceedings of the Section on Survey Research Methods, American Statistical Association*.
- WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

$$\rho_{mc} =$$

$$\sqrt{\frac{S_z^w - \frac{\sigma_z^2/\mu_z^2}{1 + \sigma_z^2/\mu_z^2} \frac{1}{n} \sum_{j \in s} w_j^2}{S_z^w - \frac{\sigma_z^2/\mu_z^2}{1 + \sigma_z^2/\mu_z^2} \frac{1}{n} \sum_{j \in s} z_j^2}}, \quad (19)$$

pour $\mu_u \neq 0$ et $\mu_v = 0$. Lorsque $\mu_u = 0$, le coefficient de $\sum w_j^2$ est $1/n$ et lorsque $\mu_v \neq 0$, le coefficient de $\sum z_j^2$ est $1/n$. Dans le modèle des questions non liées (plan à réponses randomisées (i)), l'estimateur de ρ est

$$\rho_{wq} = \frac{\sqrt{S_z^x S_z^y}}{p_1 p_2 S^{wz}}, \quad (20)$$

où $S^{uv} = N\sigma^{uv}/(N-1)$ et où

$$S_z^x = s_z^w - (1 - p_1) \frac{1}{n} \sum_{j \in s} w_j^2 + 2(1 - p_1)Uw - (1 - p_1)U^2 - (1 - p_1)\sigma_z^2 \left(p_1 + \frac{1}{1 - p_1} \right)$$

et

$$s_z^y = s_z^z - (1 - p_2) \frac{1}{n} \sum_{j \in s} z_j^2 + 2(1 - p_2)Vz - (1 - p_2)V^2 - (1 - p_2)\sigma_z^2 \left(p_2 + \frac{1}{1 - p_2} \right).$$

Lorsque $p_1 = p_2$, ce résultat peut être comparé à l'estimateur de Edgell et coll. (1986). L'estimateur résultant pour $\hat{\rho}^{uv}$ diffère de l'estimateur de Edgell et coll. (1986), lequel suppose que $\sigma^{uv} = 0$. Ces auteurs utilisent également les estimateurs biaisés de σ_z^2 et σ_y^2 . L'estimateur de Edgell et coll. pour σ_y^2 s'obtient en écrivant la variance de \bar{z} selon le plan d'expérience dans le plan d'échantillonnage aléatoire simple avec remise sous la forme suivante:

$$\sigma_z^2/n = \sum_{j=1}^N (z_j - \bar{z})^2/(Nn). \quad (21)$$

La variance de \bar{z} dans le plan à réponses randomisées est

$$[p_2\sigma_y^2 + (1 - p_2)\sigma_v^2 + p_2(1 - p_2)(Y - V)^2]/n. \quad (22)$$

L'expression (22) est tirée de Greenberg et coll. (1971). L'estimateur de σ_y^2 est calculé en l'expression (22) dans le membre de gauche de l'équation (21), en substituant l'estimateur d'échantillon de σ_z^2 et l'estimateur de la réponse randomisée de \bar{Y} dans l'équation ainsi obtenue, et en résolvant pour σ_y^2 .

5. EFFET DU BIAIS DANS LES RÉPONSES

Dans n'importe quel des plans à réponses randomisées, l'estimation la plus simple de la variance de $\hat{\rho}$ s'obtient en calculant l'estimation de la variance selon la méthode jackknife. Les estimations jackknife de la variance de $\hat{\rho}$ s'obtiennent à l'aide des formules (4.2.3) ou (4.2.5) fournies par Wolter (1985).

Chacun des estimateurs des variances et covariances de la population finie, qui sont les composantes de $\hat{\rho}$ dans (18), (19) et (20), sont non biaisés selon le plan d'expérience dans le plan à réponses randomisées approprié pour tout plan d'expérience où la probabilité de sélection composée des unités i et j donnée par $\pi_{ij} = n(n-1)/[N(N-1)]$. En conséquence, chaque estimateur est l'estimateur optimal non biaisé selon le plan d'expérience pour le paramètre de la population finie auquel il correspond. Pour obtenir les estimateurs non biaisés appropriés dans (18), on multiplie le numérateur et le dénominateur par $(N-1)/N$. Le numérateur obtenu est non biaisé selon le plan d'expérience pour σ_{xy} et les expressions placées sous le signe de la racine carrée au dénominateur de (18) sont non biaisées pour σ_z^2 et σ_y^2 . Dans (19), il est nécessaire de multiplier le numérateur et le dénominateur par $(N-1)/[N\mu_u\mu_v]$ pour obtenir la forme correcte des estimateurs non biaisés selon le plan d'expérience. Les estimateurs appropriés sont obtenus en (20) lorsque le multiplicateur est $(N-1)/(Np_1p_2)$.

Dans le modèle des constantes additives, on demande au répondant d'ajouter une variable aléatoire de u à x et une variable aléatoire indépendante de v à y . Au lieu de cela, le répondant peut décider d'ajouter des variables aléatoires indépendantes différentes, par exemple u' et v' . Les moyennes et les variances de u' et v' peuvent être différentes de celles de u et v . Il est raisonnable de penser, toutefois, que $\sigma_{u'}^2 \geq \sigma_u^2$ et $\sigma_{v'}^2 \geq \sigma_v^2$. Nous présentons ci-après un exemple d'une telle situation. Le répondant ne souhaite pas ajouter une variable aléatoire qui se situe près de la moyenne de la distribution de cette variable aléatoire. Dans ce cas, la distribution du biais dans la réponse pourrait être modélisée par la distribution originale avec un intervalle autour de la moyenne déterminé de façon que tout résultat de la distribution originale tombant dans l'intervalle choisi soit placé à l'une des extrémités de l'intervalle. En prenant séparément les espérances du numérateur et de l'expression placée sous chacun des signes de racine carrée dans le dénominateur de (18), on obtient l'expression

$$\sigma_{xy}^2 = \frac{(\sigma_x^2 + \sigma_z^2 - \sigma_{u'}^2 - \sigma_{v'}^2) \sqrt{\sigma_y^2 + \sigma_z^2} - \sigma_v^2}{\sigma_z^2}, \quad (23)$$

On peut noter dans cette expression que le biais dans les réponses conduit à une estimation de la corrélation inférieure à la valeur véritable. Le modèle des constantes multiplicatives se comporte comme le modèle des constantes additives, sauf que les réponses aux questions délicates sont multipliées par les

où \bar{z} est la moyenne d'échantillon de données et

$$s_z^2 = \frac{1}{n-1} \sum_{j \in s} (z_j - \bar{z})^2$$

est la variance d'échantillon de données obtenue par la technique à réponses randomisées, où A_2, B_2, C_2 et D_2 sont définis par les équations (7) à (10) respectivement.

Démonstration. En vertu du modèle donné par (1) et (2), la covariance $E(e_b e_c)$ est algébriquement plutôt longue, mais peut s'exprimer sous la forme

$$b^T G c + H, \quad (13)$$

où b^T est le vecteur

$$\left[E_p(b_{s..}), E_p \left(\sum_{j \in s} b_{sj} \right), E_p \left(\sum_{j \in s} b_{sll} \right) \right],$$

$$E_p \left(\sum_{i \neq j \in s} b_{sij} \right), \quad (14)$$

et c^T est un vecteur identique à (14), où tous les b sont remplacés par des c . La matrice G 4×4 en (13) contient les fonctions des moments du premier ordre de z_j et des moments du deuxième ordre de e_{zj} dans (1). Le membre

$$\sum_{i \neq j \in s} b_{sij} c_{skl}, \quad (15)$$

où le symbole de la somme est limité à un quadruplet, où les indices de b pourraient être remplacés par un point $(.)$, et où k est une fonction de moments de deuxième à quatrième ordre de e_{zj} dans (1). À noter que ces moments sont tous indépendants de j . En (15), la somme est une somme simple sur $j \in s$ lorsque, par exemple, les indices $i = j = k = l$, ou lorsque $i = k$ et que j et l sont remplacés par des points. La somme est une somme double sur $i \neq k \in s$ lorsque, par exemple, $i \neq k$ et que j et l sont remplacés par des points. Le même procédé est repris jusqu'à la quadruple somme où $i \neq j \neq k \neq l$. Selon (11), $E(e_b e_c)$ se réduit à 0 si $b_{s..} = h_1, b_{sj} = h_2, b_{sll} = h_3$, et $b_{sij} = h_4$, où les valeurs h_i sont des constantes. Selon (7) à (10) et compte tenu du fait que le plan d'expérience a une

taille fixe, nous obtenons

$$b_{s..} = D_2, b_{sj} = C_2/n, b_{sll} = -\frac{A_2 + B_2}{n(n-1)}, b_{sij} = A_2/n,$$

de sorte que l'estimateur de (12) minimise la variance de la classe non biaisée selon p et m des estimateurs quadratiques de σ_y^2 Q.E.D.

De la même manière

$$(A_1 + B_1)s_w^2 - B_1 \frac{1}{n} \sum_{j \in s} w_j^2 + C_1 w + D_1, \quad (16)$$

est l'estimateur optimal non biaisé selon p et m pour

$$A_1 = \frac{\beta_2^1(\phi_1 - \delta_1) - \delta_1 \psi_{21}}{\beta_2^1},$$

$$B_1 = \frac{\beta_2^1(\phi_1 - \delta_1) - \delta_1 \psi_{21}}{\beta_2^1 + \psi_{21}},$$

$$C_1 = \frac{(2\alpha_1 \psi_{21} - \beta_2 \psi_{11})}{\beta_2^1(\phi_1 - \delta_1) - \delta_1 \psi_{21}}, \text{ et}$$

$$D_1 = \frac{\lambda_1(\beta_2^1 + \psi_{21}) - (\alpha_1^2 \psi_{21} - \alpha_1 \beta_1 \psi_{11} + \beta_2^1 \psi_{01})}{\beta_2^1(\phi_1 - \delta_1) - \delta_1 \psi_{21}}.$$

La même méthode peut servir à estimer la covariance σ_{xy} . La classe générale des estimateurs quadratiques de σ_{xy} prend la forme de

$$e_{ds} = d_s + \sum_{j \in s} d_{1sj} z_j + \sum_{j \in s} d_{2sj} w_j + \sum_{i \neq j \in s} d_{sij} w_i z_j,$$

où les coefficients des valeurs w et z sont définis pour toutes les valeurs de s , tous les $j \in s$ et toutes les paires $(i, j) \in s$. Le résultat sur la covariance est énoncé sans

Théorème 2. En vertu du modèle défini par (1) et (2) et pour tout plan d'expérience de taille fixe n , la variance de e_d selon p et m , $E_{rp}[E_p^r(e_d - \sigma_{xy})^2] = E(e_d - \sigma_{xy})^2$, est minimisée pour l'estimateur donné par

$$s_{wz} = \frac{(\phi_3 - \lambda_3)}{(\phi_3 - \delta_3)}, \quad (17)$$

où

$$s_{wz} = \frac{1}{n-1} \sum_{j \in s} (w_j - \bar{w})(z_j - \bar{z})$$

est la covariance de l'échantillon entre w et z .

On obtient un estimateur pour ρ à partir de (12), (16) et (17). Dans le modèle des constantes additives (plan à réponses randomisées (ii)), l'estimateur de ρ est donné par

$$\hat{\rho}_{ac} = \frac{\sqrt{(s_z^w - \sigma_z^n)(s_z^v - \sigma_v^2)}}{s_{wz}}. \quad (18)$$

Il s'agit là de l'estimateur obtenu par Edgell et coll. (1986). Dans le modèle des constantes multiplicatives (plan à réponses randomisées (iii)), l'estimateur se réduit à

Dans le modèle des constantes multiplicatives, deux variables aléatoires indépendantes, u et v , assorties des moyennes μ_u et μ_v et des variances σ_u^2 et σ_v^2 respectivement, sont multipliées respectivement par la valeur de la réponse à la variable x et à la variable y . Lorsqu'on suppose un modèle de permutation aléatoire avec le couple (x_j, y_j) , on obtient dans le modèle illustré en (1) et (2)

$$\alpha_1 = \alpha_2 = 0, \beta_1 = \mu_u, \beta_2 = \mu_v,$$

$$\phi_1 = \mu_u^2 + \sigma_u^2, \phi_2 = \mu_v^2 + \sigma_v^2, \phi_3 = \mu_u \mu_v,$$

$$\psi_{21} = \sigma_u^2, \psi_{22} = \sigma_v^2,$$

$$\delta_1 = -\mu_u^2/(N-1), \delta_2 = -\mu_v^2/(N-1),$$

$$\delta_3 = -\mu_u \mu_v/(N-1), \text{ et}$$

$$\psi_{01} = \psi_{11} = \psi_{02} = \psi_{12} = \psi_3 = \lambda_1 = \lambda_2 = \lambda_3 = 0.$$

(5)

4. ESTIMATION DE LA VARIANCE ET DE LA COVARIANCE

Supposons une estimation de σ_j^2 telle que les données appropriées sont z_j pour les unités $j \in s$. La classe générale des estimateurs quadratiques de σ_j^2 prend la forme suivante:

$$e_{bs} = b_{s..} + \sum_{j \in s} b_{sj} z_j + \sum_{j \in s} b_{sjj} z_j^2 + \sum_{i \neq j \in s} b_{sij} z_i z_j, \quad (6)$$

où les coefficients des valeurs z sont définis pour tous les s , tous les $j \in s$ et toutes les $(i, j) \in s$.

Dans une situation de réponses randomisées, un estimateur e_p appartenant à la classe définie par l'équation (6) est non biaisé selon le plan d'échantillonnage pour σ_j^2 si $E_p(e_p) = \sigma_j^2$ et il est non biaisé selon le plan d'échantillonnage et le modèle de permutation (selon p et m) si $E(e_p) = \sigma_j^2$. Les conditions en vertu desquelles un estimateur e_p est non biaisé selon p et m sont réalisées en tirant l'espérance E de (6) des modèles (1) et (2). En établissant un rapport d'égalité entre les coefficients dans y_0, y_1, y_2 et σ_j^2 et on obtient quatre nouvelles équations avec quatre inconnues. La résolution de ces quatre équations conduit aux conditions suivantes en vertu desquelles les estimateurs de la classe définie par (6) sont biaisés selon p et m pour σ_j^2 :

$$E_p \left(\sum_{j \in s} b_{sij} \right) = \frac{\beta_{2j}^2 (\phi_2 - \delta_2) - \delta_2 \psi_{22}}{\beta_{2j}^2} = A_{2j}, \quad (7)$$

$$E_p \left(\sum_{i \neq j \in s} b_{sij} \right) = -\frac{\beta_{2j}^2 (\phi_2 - \delta_2) - \delta_2 \psi_{22}}{\beta_{2j}^2 + \psi_{22}} =$$

$$-(A_{2j} + B_{2j}), \quad (8)$$

$$E_p \left(\sum_{j \in s} b_{sj} \right) = \frac{(2\alpha_2 \psi_{22} - \beta_2 \psi_{12})}{\beta_2^2 (\phi_2 - \delta_2) - \delta_2 \psi_{22}} = C_2, \quad (9)$$

$$E_p(b_{s..}) = \frac{\lambda_2 (\beta_2^2 + \psi_{22}) - (\alpha_2^2 \psi_{22} - \alpha_2 \beta_2 \psi_{12} + \beta_2^2 \psi_{02})}{\beta_2^2 (\phi_2 - \delta_2) - \delta_2 \psi_{22}}$$

et

$$= D_2. \quad (10)$$

Pour obtenir l'estimateur optimal, il faut définir une classe associée d'estimateurs quadratiques de 0. Cette classe est définie par

$$e_{cs} = c_{s..} + \sum_{j \in s} c_{sj} z_j + \sum_{j \in s} c_{sjj} z_j^2 + \sum_{i \neq j \in s} c_{sij} z_i z_j.$$

Les conditions en vertu desquelles un estimateur e_c de cette classe sera biaisé selon p et m pour 0 sont définies par

$$E_p(c_{s..}) = E_p \left(\sum_{j \in s} c_{sj} \right) = E_p \left(\sum_{j \in s} c_{sij} \right) = 0. \quad (11)$$

L'obstruction de l'estimateur quadratique non biaisé selon le plan d'expérience de la variance minimale de σ_j^2 s'inspire du procédé utilisé pour la moyenne d'une population finie par Rao et Bellhouse (1978), pour les cas sans

réponse randomisée, et par Bellhouse (1980), pour les cas à réponses randomisées. La covariance $E(e_p e_c)$ pour l'espérance composite est déterminée dans le modèle de manière que $E(e_p e_c) = 0$ sous les conditions définies par (11). Les valeurs des coefficients b sont alors déterminées en vertu des conditions définies de (7) à (10). Selon un théorème de l'estimation non biaisée de la variance minimale formulé par Rao (1952), l'estimateur résultant est l'estimateur optimal non biaisé selon p et m de σ_j^2 . S'il existe un plan d'expérience tel que cet estimateur est également non biaisé selon le plan d'expérience pour σ_j^2 , on pourra considérer qu'il s'agit également de l'estimateur optimal non biaisé selon le plan d'expérience de σ_j^2 , en invoquant des arguments semblables à ceux du théorème (2.4) de Rao et Bellhouse (1978). Nous présentons d'abord les résultats obtenus pour les estimateurs non biaisés selon p et m (théorèmes 1 et 2), et ensuite les résultats obtenus pour les estimateurs non biaisés selon le plan d'expérience pour les trois processus à réponses randomisées.

Théorème 1. En vertu du modèle défini par (1) et (2) et pour les estimateurs non biaisés selon le plan d'expérience pour tout plan de taille fixe n , la variance de e_p , selon p et m , $E_{rp}[E_p(e_p) - \sigma_j^2]^2 = E(e_p - \sigma_j^2)^2$, est minimisée pour l'estimateur donné par

$$(A_2 + B_2)s_2^2 - B_2 \frac{1}{n} \sum_{j \in s} z_j^2 + C_2 \bar{z} + D_2, \quad (12)$$

paramètres des modèles de permutation aléatoire sont également des paramètres de populations finies. Dans la situation la plus simple des modèles de permutation aléatoire, on présume que le vecteur N -dimensionnel des mesures dans la population finie est une permutation aléatoire d'un vecteur N -dimensionnel de nombres fixes. Rao (1975) a montré comment cette supposition conduit à un modèle linéaire. Bellhouse (1980) a élargi l'utilisation de ce modèle aux plans à réponses randomisées à échantillonnages avec probabilités inégales.

Le modèle et les plans d'expérience apparentes applicables à l'échantillonnage avec probabilités inégales ne s'appliquent pas facilement à l'estimation des variances et des covariances, avec ou sans réponses randomisées. Nous présentons par conséquent ici un cas spécial du modèle de Bellhouse (1980). Dans ce modèle, deux opérateurs d'espérance différents interviennent et donnent ensemble une espérance composite E_m . Ces opérateurs sont: E_r , l'opérateur d'espérance lié au plan de randomisation, et E_{rp} , l'opérateur d'espérance lié au modèle de permutation aléatoire. L'espérance composite est donc $E_m = E_{rp}E_r$ et $E = E_mE_p$. Pour le modèle de permutation aléatoire, nous présumons que les paires (x_j, y_j) , $j = 1, \dots, N$ correspondent à la permutation aléatoire d'un groupe de N paires de nombres fixes que nous appellerons (p_j, q_j) , $j = 1, \dots, N$. Il s'agit là d'un cas spécial du modèle (4.1) de Rao et Bellhouse (1978); le modèle plus général décrit par Rao et Bellhouse (1978) a servi dans des cas d'échantillonnage double et d'échantillonnages effectués en deux occasions. Avec le modèle des questions non liées (plan à réponses randomisées (ii)) il faut supposer également que les quadruplets (x_j, y_j, u_j, v_j) , $j = 1, \dots, N$ constituent une permutation aléatoire d'un groupe de N quadruplets fixes de nombres que nous appellerons (p_j, q_j, r_j, t_j) , $j = 1, \dots, N$.

Supposons que la combinaison du plan de randomisation et du modèle de permutation aléatoire conduit au modèle linéaire suivant:

$$(1) \quad w_j = \alpha_1 + \beta_1 X + e_{1j} \\ z_j = \alpha_2 + \beta_2 Y + e_{2j},$$

$$\begin{aligned} \text{pour } j = 1, \dots, N \text{ où } X \text{ et } Y \text{ sont les moyennes des populations finies d'où viennent les mesures } x \text{ et } y \text{ respectivement, et où pour } j = 1, \dots, N \\ E_m(e_{1j}) = E_m(e_{2j}) = 0, \\ E_m(e_{1j}^2) = \phi_1 \sigma_x^2 + \psi_{01} + \psi_{11} X + \psi_{21} X^2, \\ E_m(e_{2j}^2) = \phi_2 \sigma_y^2 + \psi_{02} + \psi_{12} Y + \psi_{22} Y^2, \\ E_m(e_{1j} e_{2j}) = \phi_3 \sigma_{xy} + \psi_3, \text{ et} \\ E_m(e_{1j} e_{2k}) = \delta_3 \sigma_{xy} + \lambda_3, \text{ pour } j \neq k. \end{aligned} \quad (2)$$

$$(4) \quad \psi_{11} = \psi_{21} = \psi_{12} = \psi_{22} = \psi_3 = \lambda_1 = \lambda_2 = \lambda_3 = 0.$$

et tous les autres moments plus élevés sont indépendants de j . Dans le modèle illustré en (1) et (2), les paramètres α , λ , ϕ , ψ et δ sont tous des constantes connues. Les variances et les covariances des populations finies des questions délicates, σ_x^2 , σ_y^2 et σ_{xy} sont toutes inconnues. Pour le modèle des questions non liées (plan à réponses randomisées (i)), on présume que les processus de randomisation utilisés pour les deux questions délicates sont indépendants et que la question délicate i , $i = 1, 2$, est posée avec une probabilité p_i alors que les questions liées non délicates sont posées avec une probabilité $1 - p_i$. On présume en outre que les questions délicates sont sans lien avec les questions non délicates et que $\sigma_{xv} = \sigma_{yv} = \sigma_{xv} = \sigma_{yv} = 0$. Cette supposition n'est pas nécessaire dans le cas d'un échantillonnage aléatoire simple avec remise. Lorsque, en plus, on suppose un modèle de permutation aléatoire avec le quadruplet (x_j, y_j, u_j, v_j) , on obtient dans le modèle illustré en (1) et (2)

$$\begin{aligned} \alpha_1 &= (1 - p_1)U, \beta_1 = p_1, \alpha_2 = (1 - p_2)V, \beta_2 = p_2, \\ \phi_1 &= p_1, \psi_{01} = (1 - p_1)\sigma_x^2 + p_1(1 - p_1)U^2, \\ \psi_{11} &= -2p_1(1 - p_1)U, \psi_{21} = p_1(1 - p_1), \\ \phi_2 &= p_2, \psi_{02} = (1 - p_2)\sigma_y^2 + p_2(1 - p_2)V^2, \\ \psi_{12} &= -2p_2(1 - p_2)V, \psi_{22} = p_2(1 - p_2), \\ \delta_1 &= -p_1^2/(N - 1), \lambda_1 = - (1 - p_1)^2\sigma_x^2/(N - 1), \\ \delta_2 &= -p_2^2/(N - 1), \lambda_2 = - (1 - p_2)^2\sigma_y^2/(N - 1), \\ \phi_3 &= p_1 p_2, \delta_3 = -\phi_3/(N - 1), \\ \psi_3 &= (1 - p_1)(1 - p_2)\sigma_{uv}, \text{ et } \lambda_3 = -\psi_3/(N - 1). \end{aligned} \quad (3)$$

À noter qu'en raison des hypothèses de départ du modèle, il importe que la matrice des variances-covariances de la population finie des questions non délicates soit connue, ainsi d'ailleurs que les moyennes de la population finie. Pour le modèle des constantes additives (plan à réponses randomisées (iii)), on présume que les variables aléatoires u et v qui sont ajoutées à la valeur des réponses données aux deux questions délicates sont indépendantes et qu'elles sont assorties des moyennes μ_u et μ_v et des variances σ_u^2 et σ_v^2 respectivement. Lorsqu'on suppose un modèle de permutation aléatoire avec le couple (x_j, y_j) , on obtient dans le modèle illustré en (1) et (2)

$$\begin{aligned} \alpha_1 &= \mu_u, \beta_1 = 1, \alpha_2 = \mu_v, \beta_2 = 1, \\ \phi_1 &= \phi_2 = \phi_3 = 1, \psi_{01} = \sigma_u^2, \psi_{02} = \sigma_v^2, \\ \delta_1 &= \delta_2 = \delta_3 = -1/(N - 1), \end{aligned} \quad (4)$$

est désigné par E_p . Les estimateurs pour ρ sont obtenus en remplaçant σ_x^2 , σ_y^2 et σ_{xy} par leurs estimateurs respectifs, biaisés ou non, qui peuvent être optimaux dans un sens ou dans l'autre, ou ne pas l'être du tout.

Pour illustrer les résultats généraux obtenus ici dans l'estimation du coefficient de corrélation de la population finie, nous examinerons trois techniques particulières à réponses randomisées:

- (i) Le modèle des questions non liées de Greenberg et coll. (1969). La question délicate est posée avec une probabilité p , et une question sans rapport avec la première et qui n'est pas délicate est posée avec une probabilité $1 - p$. Pour l'estimation de la moyenne, on présume que la moyenne \bar{X} de la population finie de la question non liée est connue. Pour les besoins de l'estimation de la variance, on présume également que la valeur de σ_x^2 est connue.
- (ii) Le modèle des constantes additives de Pollock et Bek (1976). On ajoute une variable aléatoire issue d'une distribution de probabilité connue à la valeur de la réponse à la question délicate.
- (iii) Le modèle des constantes multiplicatives de Pollock et Bek (1976). La valeur de la réponse à la question délicate est multipliée par une variable aléatoire issue d'une distribution de probabilité connue.

Edgell et coll. (1986) ont calculé des estimateurs pour ρ dans le modèle des questions non liées et dans le modèle des constantes additives.

Dans la plupart des plans à réponses randomisées, on a présumé que le plan d'échantillonnage utilisé était l'échantillonnage aléatoire simple, avec ou sans remise. Comme les résultats dont il est question ici proviennent d'un plan d'expérience de taille fixe, on présume que l'échantillonnage aléatoire utilisé est sans remise.

Supposons que x et y sont deux variables déli-cates. On aura donc recours, pour obtenir des informations sur ces variables, à une technique à réponses randomisées. Désignons par w_j et par z_j pour $j \in s$ les mesures de l'échantillon obtenues. Désignons par u_j et par v_j pour $j = 1, \dots, N$ les mesures non déli-cates liées à x_j et y_j respectivement. En vertu du modèle de la question non liée (plan à réponses randomisées (i)), u_j et v_j sont les réponses données aux questions non liées par le j -ième répondant. En vertu du modèle des constantes additives ou du modèle des constantes multiplicatives (plans à réponses randomisées (ii) ou (iii)), u_j et v_j sont les j -ièmes variables aléatoires issues de deux distributions de probabilités connues et peut-être différentes.

3. MODELES DE PERMUTATION ALÉATOIRE

Plusieurs modèles ont été proposés pour les mesures de populations finies dans la documentation spécialisée sur l'échantillonnage d'enquête. Nous nous attarderons en particulier aux modèles de permutation aléatoire de Rao (1975) et de Rao et Bellhouse (1978). Ces modèles présentent un énorme avantage: leurs paramètres ont une interprétation directe dans la population finie étudiée puis-que les

(1988). Umesh et Peterson (1991) abordent également certaines considérations pratiques à ce propos.

Stanley Warner a contribué encore de deux façons à l'étude des réponses randomisées. La première est en rapport direct avec les résultats obtenus ici. Au milieu des multiples nouvelles idées et nouvelles techniques fondées sur le principe des réponses randomisées, Warner (1971) a formulé un modèle linéaire qui a permis d'unifier la théorie et qui permettait, à l'époque de sa formulation, d'intégrer la plupart des techniques à réponses randomisées. La seconde de ses contributions a pris en compte la tendance grandissante à recourir aux entrevues téléphoniques. Stem et Steinhorst (1984) ont décrit des méthodes à réponses randomisées applicables aux entrevues téléphoniques et aux questionnaires postaux. Warner (1986) a suggéré des plans pratiques de randomisation naturelle comme le recours aux numéros de série des billets de banque, aux fins de la réalisation d'entrevues téléphoniques.

Les principaux constituants de la méthodologie de la réponse randomisée sont: l'élaboration des techniques, la comparaison de ces techniques en tenant compte du risque de non-respect de la confidentialité du répondant, l'élaboration de plans de randomisation raisonnablement efficaces, l'élaboration d'une théorie unifiée de la réponse randomisée et la validation des techniques à l'aide d'études en conditions réelles. Stanley Warner s'est penché sur la plupart de ces grandes questions, et sa contribution en ces matières a eu une grande influence. Il est l'inventeur de la technique. Son plan d'expérience original pour une population dichotomique a été rapidement étendu au traitement des populations polymorphes et des populations soumises à des mesures en continu. Le développement des techniques à réponses randomisées s'est poursuivi et Warner a été un chef de file de l'évaluation de ces plans à l'aide de la modélisation du risque de non-respect de la confidentialité du répondant. Ses travaux en vue de l'élaboration d'un modèle linéaire unifié des plans à réponses randomisées sont devenus la pierre d'assise d'une théorie unifiée de la réponse randomisée.

2. INTRODUCTION À L'ESTIMATION DE LA CORRÉLATION

Imaginons une population finie de taille N faisant l'objet de deux mesures x_j et y_j où $j = 1, \dots, N$. Il est intéressant d'estimer la corrélation de la population finie

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y},$$

où $\sigma_{xy} = \sum (x_j - \bar{X})(y_j - \bar{Y})/N$ représente la covariance de la population finie entre les variables x et y , et où σ_x^2 et σ_y^2 sont les variances des variables x et y respectivement. Pour faire l'estimation de ρ , on choisit au sein de la population finie un échantillon de taille définie n assorti d'une probabilité $P(s)$, où s désigne le groupe d'unités de la population finie choisi pour l'échantillon. L'opérateur de l'éspérance rattaché à la probabilité $P(s)$ du plan d'échantillonnage

Estimation de la corrélation dans les plans à réponses randomisées

D.R. BELLHOUSE¹

RÉSUMÉ

Nous examinons la contribution de Stanley Warner à l'étude des réponses randomisées, avant d'élaborer un modèle linéaire, fondé sur les modèles de permutation aléatoire, qui permet d'intégrer, à titre de cas spéciaux, plusieurs plans connus à réponses randomisées. On obtient avec l'aide de ce modèle des estimateurs optimaux des covariances et des covariances de populations finies à l'intérieur d'une classe générale d'estimateurs quadratiques non biaisés. Ces résultats conduisent à la mise au point d'un estimateur de la corrélation de la population finie. Trois plans d'expérience particuliers à réponses randomisées sont examinés: (i) le modèle des questions non liées de Greenberg et coll. (1969); (ii) le modèle des constantes additives de Pollock et Bek (1976); le modèle des constantes multiplicatives de Pollock et Bek (1976). Nous proposons des modèles simples du biais dans les réponses afin d'illustrer l'effet de ce biais sur l'estimation de la corrélation.

MOTS CLÉS: Modèle des constantes additives; modèles linéaires; modèle des constantes multiplicatives; biais dans les réponses; modèle des questions non liées; estimation de la variance.

1. APERÇU DE LA CONTRIBUTION DE WARNER À L'ÉTUDE DES RÉPONSES RANDOMISÉES

On a recours à la technique des réponses randomisées pour obtenir des réponses à des questions délicates. Cette méthode a été mise au point il y a trente ans par Stanley Warner (Warner 1965) afin d'estimer une proportion avec un plan d'échantillonnage aléatoire simple avec remise. Ce progrès théorique remarquable a été rendu possible grâce à une démarche intellectuelle d'une grande originalité. Comment obtenir des réponses sincères à des questions délicates? Warner a proposé comme solution d'obtenir ces réponses sans que l'intervieweur n'ait conscience que les questions posées sont réellement délicates. Il a élaboré la structure probabiliste de l'interrogation de manière à obtenir une estimation de la proportion requise. Dans la formule originale de Warner, la population est divisée en deux groupes exhaustifs mutuellement exclusifs: A et B. Il importe d'estimer la proportion π de la population appartenant au groupe A. Pour y arriver, on utilise une simple aiguille pivotante qui pointera en direction d'une lettre A avec une probabilité p , et en direction d'une lettre B avec une probabilité $1 - p$. La personne questionnée fait tourner l'aiguille et n'a qu'à répondre oui ou non selon que cette aiguille s'arrête ou non sur son groupe d'appartenance. Le plan d'échantillonnage avec remise permet d'estimer π par la méthode du maximum de vraisemblance. Cette idée d'une grande originalité a beaucoup attiré l'attention au cours des trente dernières années. Depuis les travaux originaux de Warner, plusieurs techniques à réponses randomisées ont été proposées pour l'estimation d'une proportion ou d'un groupe de proportions – par exemple, les données polytomiques – ou pour l'estimation de la moyenne d'une population à partir de données continues. Une variante de l'idée originale de Warner,

proposée pour la première fois par Greenberg et coll. (1969), consiste à poser la question délicate ou une question non liée avec des probabilités de p et de $1 - p$ respectivement. D'autres variantes possibles utilisant des données continues comprennent l'ajout d'une variable aléatoire à la réponse donnée à la question délicate, ou la multiplication de la réponse par une variable aléatoire. L'idée qui sous-tend toutes ces techniques est de masquer la réponse originale de manière qu'il devienne impossible d'attribuer l'information obtenue à l'un ou l'autre des répondants, tout en permettant l'extraction de l'information portant sur la question délicate à partir de l'échantillon général. Il existe aujourd'hui une abondante documentation sur ces techniques, y compris une monographie de Chaudhuri et Mukerjee (1988). Nathan (1988) a préparé une liste bibliographique assez complète de tous ces ouvrages. Umesb et Peterson (1991) ont pour leur part fourni plusieurs exemples à réponses randomisées.

Compte tenu de la variété actuelle des techniques à réponses randomisées, on peut s'interroger sur la façon de les comparer. La minimisation de la variance n'est pas le seul critère à prendre en compte. Chaque méthode est conçue pour protéger la vie privée du répondant. Un gain d'efficacité, en termes de variance, obtenu grâce au choix de valeurs différentes de la probabilité dans le plan de randomisation ou au choix d'une méthode à réponses randomisées au détriment d'une autre, peut mettre en péril le caractère confidentiel des réponses fournies. Pour répondre à cette préoccupation, Leysieffer et Warner (1976) et Warner (1976) ont proposé des mesures naturelles du risque de non-respect de la confidentialité des répondants. Ces mesures sont liées à la probabilité, pour un intervieweur, de déterminer l'origine d'une réponse à une question délicate. La théorie du risque de non-respect de la confidentialité a été examinée par Chaudhuri et Mukerjee

¹ D.R. Bellhouse, Department of Statistical and Actuarial Sciences, University of Western Ontario (London) Ontario, N6A 5B7.

BIBLIOGRAPHIE

- BISHOP, Y.M.M., FIENBERG, S.E., et HOLLAND, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge: MIT Press.
- FIENBERG, S.E., et TANUR, J.M. (1988). From the inside out and outside in: Combining experimental and sampling structures. *La Revue Canadienne de Statistique*, 16, 135-151.
- FIENBERG, S.E., (Ed.) (1989). *The Evolving Role of Statistical Assessments as Evidence in the Courts*. New York: Springer-Verlag.
- GOOD, I.J. (1950). *Probability and the Weighing of Evidence*. London: Griffin.
- NELDER, J.A., et WEDDERBURN, R.W.M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, A*, 135, 370-384.
- WARNER, S.L. (1975). Advocate scoring for unbiased information. *Journal of the American Statistical Association*, 70, 15-22.

- WARNER, S.L. (1979). Subjective information in statistics. *Proceedings of the Business and Economics Section, American Statistical Association*, 558-563.
- WARNER, S.L. (1981). Balanced information, The Pickering Airport experiment. *The Review of Economics and Statistics*, LXII, 256-262.
- WARNER, S.L. (1984). An overlapping information survey model for evaluating summary information. *Proceedings of the Social Statistics Section, American Statistical Association*, 581-584.
- WARNER, S.L. (1985). Applications of the overlapping information model. *Proceedings of the Section on Survey Research Methods, American Statistical Society*, 401-403.
- WARNER, S.L. (1987a). Identifying rational opinion-formation with the overlapping information model. In *Applied Probability, Stochastic Processes, and Sampling Theory*. (Eds. I.B. MacNeill et G.J. Umphrey). Dordrecht, The Netherlands: Reidel, 323-329.
- WARNER, S.L. (1987b). Using test populations to develop balanced agenda. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 441-443.
- WARNER, S.L. (1992). Statistically Balanced Information Technology. Manuscript non-publié.

Warner (1985, 1987a) est revenu sur ce thème de l'information chevauchante, et il a étendu le modèle de l'expression (14) à la forme suivante:

$$I_i = mS'_i + D'_i r_i (Z_i - U_i), \tag{16}$$

où, en fait, le coefficient $-m/N$ de l'expression (14) a été remplacé par $D'_i r_i$, où $D_i \geq -1$ est un facteur de dévaluation et $E(r_i) = m/N$. Warner a ensuite montré comment estimer les coefficients de ce modèle de régression à "coefficients aléatoires" au moyen des moindres carrés généralisés, selon diverses hypothèses quant aux corrélations entre les quantités de (16).

Il a ensuite appliqué cette méthode à un nouvel ensemble de données provenant d'une enquête téléphonique auprès d'étudiants de l'Université Carleton sur la pertinence d'avoir un Sénat canadien élu plutôt que de maintenir le système de nomination des sénateurs. Les répondants ont été invités à se prononcer sur le sujet sous forme d'une probabilité. On leur a ensuite présenté un sommaire de six phrases d'un débat télévisé ayant eu lieu sur le sujet et on leur a demandé de réévaluer la probabilité qu'ils avaient indiquée. Des 417 participants, 316 avaient donné au départ des probabilités différentes de 0 et 1. De ce groupe, 163 ont modifié leur évaluation et, dans l'ensemble, la cote logarithmique moyenne après la présentation du sommaire était pratiquement la même qu'avant, mais avec une variance légèrement plus faible. Warner a ajusté le modèle aux données, et les équations ajustées étaient compatibles avec la notion d'une dévaluation partielle de l'information dont les participants avaient déjà eu connaissance.

Ce fut essentiellement la dernière publication de Warner sur le problème de l'évaluation d'une information équilibrée. Au moment de son décès, Warner travaillait intensément sur le manuscrit d'un livre dont le titre, *Statistically Balanced Information Technology*, laisse croire qu'il tentait de synthétiser et de développer ses idées sur le sujet. Malheureusement, nous n'avons pu trouver que les premiers chapitres de ce livre, qui contiennent seulement les idées de base des méthodes de calcul de probabilité et de régression qu'il comptait utiliser dans les chapitres ultérieurs.

5. OBSERVATIONS COMPLÉMENTAIRES

La technologie de l'information équilibrée de Warner s'attaque au problème courant des politiques controversées, qui peut être la source de confusion et de mauvaises décisions en raison du déséquilibre dans la présentation des faits pertinents. Des exemples de la façon dont une approche contradictoire à la résolution de conflits dans un contexte juridique pourrait avoir des effets de distorsion sur des sujets scientifiques sont présentés dans Fienberg (1989; voir en particulier l'annexe H par Vidmar). Parmi les solutions fréquemment proposées pour résoudre ce problème figure la création d'un tribunal scientifique qui veillerait à l'équilibre de l'information relative à un différend factuel et servant à la prise de décisions. Le tribunal

scientifique ainsi proposé est un système contradictoire, mais fondé sur des règles bien définies pour la sélection des sujets, des porte-parole ainsi que des juges désignés pour veiller à l'impartialité et atténuer le plus possible l'impact des préférences personnelles. L'approche de Warner décrite ici est un moyen formel conçu exactement pour atteindre ce genre d'impartialité.

L'évolution des travaux de Warner sur l'information équilibrée s'est accompagnée d'un changement de sa perception des fondements des statistiques. Il a reçu une formation d'économiste et de statisticien classique, et ses premières contributions statistiques, notamment ses travaux sur les modèles à réponses randomisées, s'appuyaient sur des bases statistiques dénuées de subjectivisme. Son article de 1975 sur la notation des partisans d'une position a été le premier indice d'une démarche subjectiviste et, dans chaque article subséquent, il a intégré de nouveaux éléments de l'approche bayésienne. Dans Warner (1979), il révèle clairement l'évolution de sa pensée à ce sujet, et le changement apparaît clairement dans les premiers chapitres de son livre non publié. À l'assemblée annuelle de mai 1992 de la Société statistique du Canada à Edmonton, lors de sa dernière conférence publique, Warner a décrit les méthodes d'obtention de probabilités qu'il avait élaborées pour son livre.

Nous ne pouvons qu'échafauder des hypothèses sur la forme que Warner aurait donnée à sa synthèse subjectiviste de la technologie de l'information équilibrée s'il avait pu terminer son livre. Toutefois, compte tenu de son adhésion profonde à l'approche bayésienne et des récentes innovations méthodologiques de cette approche, nous croyons que sa méthode aurait inclus un modèle linéaire généralisé hiérarchique et utilisé les derniers développements des techniques de simulation de Monte Carlo à chaînes de Markov.

Stanley Warner utilisait constamment dans ses cours les idées découlant de ses recherches. Evoquant les travaux que nous venons de décrire, il disait:

"... presque tous les éléments d'un cours de statistique élémentaire se retrouvent à un degré ou à un autre dans ces méthodes, et les problèmes soulevés sur le plan de la modélisation et de la conception pourraient être examinés à un niveau très avancé," (Warner 1987b) (Traduction).

REMERCIEMENTS

La préparation du présent article a été rendue possible en partie par une subvention du Conseil de recherches en sciences naturelles et en génie du Canada, accordée au premier auteur, à l'Université York.

Chaque analyse s'est fondée sur l'utilisation d'une forme différente de moindres carrés pondérés pour l'estimation des coefficients d'intérêt.

Tableau 4

Estimations de l'influence des arguments pour l'expérience sur le projet de l'aéroport de Pickering

Paramètres	Influence globale	Influence désagrégée	Influence globale pondérée	Influence moyenne
β_1	-.857 (.093)	-.529 (.047)	-.736 (.065)	
β_2	.485 (.188)	.337 (.097)	.462 (.132)	
β_3	.147 (.188)	.146 (.097)	.187 (.132)	
β_4	.313 (.186)	.209 (.095)	.307 (.129)	
N	8	8	8	567

Source: Warner (1981).

Nous présentons au tableau 4 les coefficients estimés de Warner pour les trois modèles. Tous les modèles donnent des résultats semblables, et les résultats peuvent être résumés comme suit:

a) C'est l'équipe 2 qui, manifestement, a présenté l'argumentation la plus solide (β_2 et β_3 sont tous deux positifs et semblables dans les trois colonnes).

b) L'influence globale estimée pour l'équipe 2 est de $[\beta_1 + .5(\beta_2 - \beta_3)] = -0.688$, ce qui correspond à une proportion estimée de répondants en faveur du projet d'aéroport de $\beta = 0.355$.

c) L'influence désagrégée pour l'équipe 2 correspond à une proportion estimée de répondants en faveur du projet d'aéroport de $\beta = 0.355$.

d) L'effet de l'ordre de présentation (β_4) laisse supposer que l'exposé figurant en premier dans l'envoi a eu le plus grand impact, ce qui est conforme à l'hypothèse selon laquelle "une information antérieure favorable à une position contribue à dévaluer une nouvelle information qui s'oppose à cette position".

Il appert que le conseiller de l'équipe 1 jugeait que la construction de l'aéroport n'était pas justifiable, ce qui a sérieusement hypothéqué les efforts de l'équipe 1 favorable au projet (ce que reflètent les estimations de β_2).

4. INFORMATION CHEVAUCHANTE

Warner s'inquiétait du fait que l'information utilisée dans l'expérience sur l'aéroport de Pickering comportait un chevauchement des arguments "pour" et "contre", et qu'il y avait aussi un chevauchement de l'information préalable dont disposaient les répondants et de celle pré-sentée par les partisans des deux positions. Il s'est penché

est alors éléments d'information déjà vus. L'information ajoutée "rationnelle" et ne se laissent plus influencer par des présents. Supposons que les participants agissent de façon présentées utilisent un sous-ensemble, S , de m des N éléments d'information. On peut alors considérer que l'information sans remise. Les arguments sommaires "pour" ou "contre" sans remise à même le total N , nous pouvons considérer que les éléments de $A(i) \cap S$ ont été sélectionnés au hasard.

Si les m éléments d'information sont choisis au hasard sans remise à même le total N , nous pouvons considérer que les éléments de $A(i) \cap S$ ont été sélectionnés au hasard sans remise. Les arguments sommaires "pour" ou "contre" sans remise à même le total N , nous pouvons considérer que les éléments de $A(i) \cap S$ ont été sélectionnés au hasard.

$$I_i = mS_i - m/N[p_i/(1 - p_i)] + \epsilon, \quad (14)$$

où I_i est l'information nette du sommaire pour la i -ième personne. S_i est la moyenne des Y_{ij} pour les éléments d'information $j \in S$, et le terme d'erreur ϵ a une espérance conditionnelle nulle, c.-à-d.:

$$E\{ \epsilon_i | [p_i/(1 - p_i)] \} = 0. \quad (15)$$

huit du plan $2 \times 2 \times 2$ de la section 2, plus les deux échantillons de contrôle formés d'économistes à qui l'on avait demandé de se prononcer sans lire d'exposés. Un total de 726 économistes ont participé à l'expérience. Nous présentons au tableau 3 les sommaires des résultats de Warner pour les huit sous-échantillons expérimentaux; les opinions "après exposés" ont été classées dans trois groupes selon que le résultat était sensiblement supérieur, à peu près égal ou sensiblement inférieur à 0.5; elles ont en outre été regroupées sur l'ensemble des huit sous-échantillons expérimentaux. Les résultats ont été stratifiés a posteriori selon que les économistes étaient professeurs, étudiants des cycles supérieurs ou autres. Les données "avant exposés" reflètent les résultats combinés des deux groupes de contrôle.

Tableau 3

Opinions de la population cible sur le projet de l'aéroport de Pickering

Professeurs			Étudiants			Autres			Totaux		
Avant exposés	Après exposés	Contrôle	Avant exposés	Après exposés	Contrôle	Avant exposés	Après exposés	Contrôle	Avant exposés	Après exposés	Contrôle
9	58	32	9	32	11	71	29	161	19	567	567
(.143)	(.266)	(.184)	(.257)	(.288)	(.180)	(.298)	(.182)	(.284)	(.033)	(.033)	(.033)
32	155	72	12	72	36	160	80	387	22	387	387
(.508)	(.711)	(.343)	(.648)	(.590)	(.672)	(.503)	(.683)	(.683)	(.349)	(.349)	(.349)
22	5	7	14	7	14	7	50	19	63	218	218
(.349)	(.023)	(.063)	(.400)	(.230)	(.029)	(.315)	(.033)	(.033)	(1.000)	(1.000)	(1.000)
63	218	111	35	111	61	238	159	567			
(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)			

Source: Warner (1981).

Notons que les trois groupes étaient, après les exposés, nettement défavorable à l'aéroport proposé et que les écarts entre les proportions d'indécis des sous-échantillons expérimentaux et celles des groupes de contrôle révèlent que les exposés des positions ont influé sur les opinions. L'analyse statistique formelle des données faite par Warner s'est concentrée seulement sur les huit sous-échantillons expérimentaux et s'est fondée sur trois variantes du modèle formel de l'expression (9) ainsi que sur une redéfinition des paramètres des expressions (10) à (13):

- i) Une structure logit semblable à celle de Warner (1975) et fondée sur les regroupements du tableau 3, dans laquelle, en fait, les "indécis" sont *imputés* à la catégorie "pour" ou à la catégorie "contre" avec une probabilité de 0.5. Il s'agit, selon la désignation de Warner, d'un modèle de l'influence globale simple.
- ii) Une approche plus directe, consistant à faire la moyenne des évaluations après exposés pour obtenir des "proportions globales" favorable au projet, puis à traiter ces proportions observées comme si elles étaient binomiales. Warner la désignait comme un modèle de l'influence globale pondérée.
- iii) Un modèle à deux degrés, dans lequel les évaluations de 17 niveaux individuels de l'intervalle 0-1 étaient utilisées, après quoi un modèle de régression "à coefficients variables" était analysé. Warner l'a baptisé modèle de l'influence désagrégée moyenne.

Il convient de souligner que la première version de cet article a été soumise pour publication au *JASA* en juin 1992, avant que Warner ait effectué l'étude empirique sur la controverse de la voie rapide Spadina. Plus de deux ans se sont écoulés avant qu'il ne soumette une version révisée de l'article comportant l'exemple détaillé. Même les auteurs connus et aux idées innovatrices ont souvent de la difficulté à faire publier leurs travaux dans les grandes publications statistiques, et une application empirique aux résultats non équivoques est toujours utile.

3. DÉVELOPPEMENT DE LA MÉTHODE ET DEUXIÈME APPLICATION

Warner a développé son approche en matière d'information équilibrée dans un deuxième article (Warner 1981), qui présentait une autre application. Cet article marque aussi une importante évolution de la pensée de Warner sur la statistique et les probabilités: détaillant les approches classiques qu'il favorisait au début de sa carrière, il tendait vers une approche subjective de type bayésien. Bien que les analyses dont il fait état demeurent axées sur l'observation de fréquences, Warner a utilisé, au moins de manière informelle, les évaluations de probabilités antérieures d'une manière reliée assez naturellement à la formulation bayésienne de l'expression (1) ci-dessus. En mars 1972, le gouvernement fédéral canadien a annoncé un projet de construction d'un deuxième aéroport international à Toronto, à l'est de la ville de Pickering, en Ontario. Ce projet a suscité une vaste controverse. En 1974, le gouvernement a mis sur pied une commission d'enquête de trois personnes. Warner a effectué une expérience simul-tanée, mais indépendante. Il a demandé aux répondants d'indiquer si oui ou non l'aéroport de Pickering devrait être construit avant l'an 2000. La structure générale de l'expérience était semblable à celle de l'expérience précédente sur la voie rapide Spadina, mais avec certaines différences: i) La population étudiée était cette fois composée d'écono-

nomistes. ii) Il a incorporé deux sous-échantillons de contrôle "neutres", auxquels ni les arguments "pour" ni les arguments "contre" n'ont été présentés. iii) Les répondants des 8 sous-échantillons expérimentaux ont donné des évaluations de probabilités (plutôt que des valeurs 0-1) après avoir évalué les positions défendues. Les répondants des groupes de contrôle ont aussi donné leurs évaluations de probabilités. La population de l'expérience se limitait aux économistes membres de la Canadian Economic Association, ou encore à ceux qui étaient professeurs ou chargés de cours en sciences économiques dans une université canadienne. L'enquête comportait deux envois postaux. Le premier a permis de déterminer quels économistes étaient disposés à lire des exposés détaillés et à "faire part de leur opinion sur un projet fédéral non divulgué". Pour le deuxième envoi postal, les économistes qui avaient accepté de participer ont été divisés en dix sous-échantillons, soit les

$$\ln[p_{ijk}/(1 - p_{ijk})] = \ln[P(H)/P(\bar{H})]$$

$$+ I(F_i) - I(A_j) + D_k. \quad (7)$$

L'expression (7) apparaît alors comme un modèle logit linéaire, et le plan d'échantillonnage comme un plan produit-binomial (sans tenir compte de la correction associée à la fraction de sondage de 0,2%). Certes, ces travaux de Warner sont antérieurs à la publication d'une monographie de Bishop, Fienberg et Holland (1975) et remontent pratiquement à la même date que l'article de Nelder et Wedderburn (1972) sur les modèles linéaires généralisés. Son article ne faisait par conséquent aucune allusion aux modèles logit ou log-linéaires, qui ont fait l'objet de multiples publications depuis ce temps.

Tableau 1

Préférences, selon les échantillons, à l'égard de la voie rapide Spadina après information par des partisans des positions

Echantillon	i	j	k	Pour	Contre	Indécis	Total	p_{ijk}	n_{ijk}
1	1	1	1	22	4	1	27	.846	26
2	1	1	2	18	9	2	29	.666	27
3	1	2	1	26	8	0	34	.764	34
4	1	2	2	21	11	1	33	.656	32
5	2	1	1	28	10	1	39	.736	38
6	2	1	2	14	11	1	26	.560	25
7	2	2	1	19	16	1	36	.542	35
8	2	2	2	19	17	2	38	.527	36

Source: Warner (1975).

Pour estimer les paramètres de l'expression (7), Warner a utilisé les moindres carrés pondérés, qui donnent à la fois les coefficients estimés et les erreurs types. Plutôt que de traiter directement les paramètres de l'expression (7), il les a redéfinis, d'une part pour simplifier les calculs et d'autre part pour faciliter leur interprétation.

$$\beta_1 = \ln \frac{P(H)}{P(\bar{H})} + \frac{I(F_1) - I(A_1)}{2} + \frac{I(F_2) - I(A_2)}{2}$$

$$+ \frac{D_1 + D_2}{2}, \quad (8)$$

$$\beta_2 = I(F_1) - I(F_2), \quad (9)$$

$$\beta_3 = I(A_1) - I(A_2), \quad (10)$$

$$\beta_4 = D_1 - D_2. \quad (11)$$

Le coefficient β_1 est un paramètre de "coordonnée à l'origine" ou de normalisation, tandis que β_2 , β_3 et β_4 mesurent la performance des équipes de partisans, et que β_4 mesure l'effet de l'ordre. L'information nette fournie par l'équipe 1 est $\beta_1 + .5(\beta_2 - \beta_3)$, et celle fournie par l'équipe 2 est $\beta_1 - .5(\beta_2 - \beta_3)$. La différence d'influence nette est donc $\beta_2 - \beta_3$.

Tableau 2
Estimations des paramètres théoriques selon les moindres carrés pondérés

Paramètre	Estimation	Erreur type approx.
β_1	.712	.139
β_2	.648	.277
β_3	-.383	.275
β_4	.528	.274
$\beta_2 + \beta_3$.264	.386
$\beta_1 + .5\beta_2 - .5\beta_3$	1.228	.266
$\beta_1 - .5\beta_2 + .5\beta_3$.196	.215
$\beta_2 - \beta_3$	1.032	.395

Source: Warner (1975).

Nous reproduisons les résultats des estimations de Warner du tableau 2. Nous avons vérifié les valeurs estimées du tableau 2 au moyen des programmes de modèles généralisés de S^+ , qui utilisent une version de la méthode des moindres carrés pondérés itérativement (maximum de vraisemblance dans ce cas). Les résultats de notre modèle logit concordent avec ceux de Warner jusqu'à deux décimales. La somme des carrés des écarts résiduels de ce modèle est de 1.95 avec 4 degrés de liberté, ce qui révèle une qualité d'ajustement remarquable et donne une forte crédibilité à l'hypothèse de l'indépendance de l'expression (4).

Dans son interprétation des résultats du tableau 2, Warner a indiqué que son analyse économique l'amenait à conclure que la proportion globale de la population favorable à H , en présence d'une information non biaisée, se situe entre les estimations "pures" pour les deux équipes de partisans, c.-à-d., dans le présent cas, dans l'intervalle (0.55, 0.77). Ces limites correspondent aux estimations l'indiquait clairement le tableau 1, peu importe comment les arguments "pour" et "contre" étaient combinés, la majorité des personnes de chaque sous-groupe favorisait le parachèvement de la voie rapide. Warner a signalé qu'on pourrait être tenté d'utiliser β_1 pour produire une "meilleure estimation" de la valeur de p correspondant à une information non biaisée, mais il plaiderait pour une valeur supérieure, puisque l'équipe 1 est supérieure à l'équipe 2 en termes d'information totale, c.-à-d. que $\beta_2 - \beta_3 > 0$. (La supériorité de l'équipe 1 ressort clairement d'un examen rapide du tableau 1, sans qu'on ait besoin de l'analyse complète.)

Warner terminait son article de 1975 en signalant toutes les lacunes de son expérience de faible portée et de ses efforts de modélisation initiaux. Ce qui retient l'attention en rétrospective, c'est sa capacité de s'attaquer à un problème très complexe de politique publique et d'enquête au moyen d'un modèle simple mais ingénieux, de construire un plan d'estimation rigoureux et solide fondé sur l'incorporation d'une expérience à une enquête par sondage et d'appliquer la méthode pour trouver une réponse à un problème concret.

Pour s'assurer que les partisans des positions traitent équitablement aussi bien la position "pour" que la position "contre", Warner a proposé de les répartir en fonction de la somme d'information nette qu'ils apportaient, c.-à-d.

$$(5) \quad I(F_i) + I(A_j).$$

La théorie économique, selon Warner, permet de croire que le fait de répartir les partisans de cette façon les à tenter, au moins, d'approcher le niveau de l'information non biaisée" associée à la maximisation avec contrainte de ressources. Nous devons par conséquent estimer la quantité de l'expression (5) ainsi que la cote a posteriori fondée sur une information non biaisée:

$$(6) \quad P(H | F', A') | P(H | F', A').$$

L'«équilibre» dans la méthode de collecte des données était la clé du plan d'estimation de Warner.

Le plan d'estimation de Warner était lié à l'application étudiée. La question controversée était le parachèvement de la voie rapide Spadina dans l'axe nord-sud de Toronto (où résidait Warner). Un tronçon initial de la voie rapide avait été construit en 1966 et, après de nombreux débats, le reste du projet avait été annulé en 1971. Deux ans plus tard, en 1973, Warner a effectué une enquête pour connaître quelles proportions de la population des électeurs inscrits de la région métropolitaine de Toronto étaient pour et contre le projet initial d'autoroute. Il a prélevé un échantillon aléatoire de 1,360 électeurs inscrits (1% de la population visée), qu'il a divisé en 8 sous-échantillons égaux de 170 personnes. Deux équipes de partisans ont préparé chacune une position écrite "pour" la voie rapide et une position écrite "contre" la voie rapide, et chaque envoi postal contenait un énoncé de position "pour" et un autre "contre". On a également fait varier l'ordre de présentation des deux positions écrites. Il s'agissait donc d'un plan d'expérience $2 \times 2 \times 2$, dans lequel la première variable correspondait à l'auteur de la position "pour", la deuxième, à l'auteur de la position "contre", et la troisième, à l'ordre de présentation ("pour" en premier ou "contre" en premier). Les partisans ont reçu une rémunération de base, et un montant supérieur a été réservé à l'équipe obtenant la "meilleure note combinée". Il s'agit d'un excellent exemple d'une expérience factorielle incorporée à une enquête, conforme à l'esprit des expériences incorporées selon la description de Fienberg et Tanur (1988).

Dans la lettre d'accompagnement, Warner demandait aux répondants de retourner des cartes-réponse pré-affranchies indiquant leur préférence après avoir examiné les positions. À la date limite, 262 cartes avaient été reçues, pour un taux de réponse d'environ 20%. Les résultats, tirés de l'article de Warner (1975), sont présentés au tableau 1. Soit p_{ijk} la proportion réelle de la population qui est "pour" la voie rapide dans le groupe (i, j, k) . Alors, avec un terme additif reflétant l'ordre de présentation des positions, le modèle de l'expression (1) devient

2. Comment peut-on noter les partisans des positions dans de telles situations?

Warner a élaboré une formulation visant à répondre en même temps aux deux questions et, ce faisant, il a utilisé à la fois des arguments économiques et des arguments statistiques. Dans le présent article, nous nous concentrons sur le volet statistique de son argumentation et nous renvoyons à l'article de Warner les lecteurs intéressés au volet économique.

Considérons deux partisans, ou deux équipes de partisans, dont le rôle est d'exposer à un auditoire leurs arguments sur une question controversée, H . Désignons par $P(H)$ et $P(H)$ les proportions du nombre de personnes d'une population donnée qui sont "pour" et "contre" la question H . Appelons F_i et A_j les présentations "pour" et "contre" des partisans i et j , respectivement, pour $i, j = 1, 2$. Soient $P(H | F_i, A_j)$ et $P(H | F_i, A_j)$ les nombres de personnes "pour" et "contre" la question H après avoir entendu les arguments "pour" du partisan i et les arguments "contre" du partisan j .

Warner a défini l'«information nette» associée à F_i et à A_j comme suit:

$$I(F_i, A_j) = \ln[P(H | F_i, A_j) / P(H | F_i, A_j)] - \ln[P(H) / P(H)]. \quad (1)$$

La formule (1) est, évidemment, le logarithme du facteur de Bayes, ou ce que Good (1950) appelait le *poids de la preuve*. Bien que Warner ait reconnu la nature évocatrice de l'utilisation du théorème de Bayes dans ce cas, son approche s'intéressait purement à l'observation de fréquences. De même, Warner a défini séparément l'information nette associée à l'argumentation "pour" et "contre" de F_i et A_j :

$$I(F_i) = \ln[P(H | F_i) / P(H)] - \ln[P(H) / P(H)], \quad (2)$$

$$I(A_j) = \ln[P(H | A_j) / P(H)] - \ln[P(H) / P(H)]. \quad (3)$$

L'hypothèse la plus simple qu'on puisse faire au sujet des quantités conjointes et marginales d'information est celle de l'indépendance des argumentations "pour" et "contre",

$$(4) \quad I(F_i, A_j) = I(F_i) + I(A_j),$$

pour $i = 1, 2$ et $j = 1, 2$. Cette hypothèse permet certaines comparaisons directes et, comme nous le verrons, peut être vérifiée empiriquement.

Contributions de Stanley Warner à la technologie de l'information statistiquement équilibrée

STEPHEN E. FIENBERG et NURI JAZAIRI¹

RÉSUMÉ

Stanley Warner s'est acquis une renommée à titre de concepteur de la technique de randomisation des réponses pour les enquêtes comportant des questions délicates. Pendant près de vingt ans, il a aussi formulé et mis au point une méthode statistique s'appliquant à un autre problème, celui de l'obtention d'une information équilibrée sur des questions controversées de telle sorte que les deux points de vue soient représentés de manière adéquate et équilibrée. Nous passons en revue ces travaux, notamment deux situations d'enquête dans lesquelles Warner a mis sa méthode à l'épreuve, et nous examinons comment les concepts proposés s'inscrivent dans l'évolution actuelle de la méthodologie.

MOTS CLÉS: Notation de partisans; théorème de Bayes; expérience incorporée; régression logistique; analyse d'enquête.

1. INTRODUCTION

Voici des exemples récents de questions d'ordre public ou professionnel suscitant la controverse:

1. Le Canada devrait-il adhérer à l'Accord de libre-échange nord-américain?
2. Le Québec devrait-il se séparer de la fédération canadienne?
3. L'American Statistical Association devrait-elle adopter un programme de certification des statisticiens?
4. L'usage du tabac devrait-il être banni dans tous les restaurants d'Ottawa?

Les débats qui entourent ce genre de questions reflètent souvent des points de vue nettement contradictoires, et les arguments "pour" et "contre" peuvent contribuer largement à façonner l'opinion des personnes appartenant aux populations étudiées (p. ex. les citoyens canadiens, les membres de l'ASA, les clients des restaurants d'Ottawa). Le présent article s'intéresse à la possibilité de présenter ces points de vue contradictoires d'une façon équilibrée. Il a souvent été dit que seule une faible proportion des scientifiques font, une fois dans leur vie, une contribution réellement novatrice à la recherche. Un nombre encore beaucoup moindre ont à leur crédit plusieurs innovations. Stanley Warner s'est fait connaître par la création et la mise au point du modèle des réponses randomisées pour les enquêtes, contribution qui a été reconnue comme un apport majeur à la statistique. Mais on lui doit aussi une approche moins bien connue mais néanmoins innovatrice face au problème de l'information équilibrée dans des contextes de controverse, sur laquelle il a travaillé pendant près de vingt ans. À titre de collègues de Stanley Warner à l'Université York au moment de son décès en 1992, nous savons à quel point il considérait comme une obligation de la statistique et des statisticiens de se pencher sur des

2. LE PROBLÈME DE BASE

Dans un article à la pensée riche, bâti sur une argumentation solide et suscitant la réflexion, publié en 1975 dans le *Journal of American Statistical Association*, Stanley Warner a soulevé pour la première fois le problème consistant à mesurer l'impact sur l'opinion publique des positions exprimées et de l'équilibre de l'information dans des contextes de controverse. Pour cela, il a posé deux questions interdépendantes (et y a ensuite répondu):

1. Comment pouvons-nous estimer ce que conclurait la population sur une question si chaque personne appartenant à cette population recevait une information équilibrée sur cette question?

Warner a eu l'idée, pour répondre à cette question, de demander à des partisans des positions en cause de

Le but du présent article est de donner aux membres de la profession une présentation neuve des idées de Warner sur le problème de l'information équilibrée dans des contextes de controverse, et de montrer comment ces idées s'intègrent aux pratiques et aux tendances méthodologiques actuelles dans le domaine des enquêtes. À la section 2, nous présentons son approche de base du problème et nous décrivons le modèle statistique qu'il a privilégié (Warner 1975). Aux sections 3 et 4, nous examinons des améliorations de son approche de base qu'il a présentées dans des articles subséquents (p. ex. voir Warner 1981, 1984, 1985, 1987a), puis nous décrivons comment Warner a poursuivi ce programme de recherche jusqu'à son décès. Tout au long de cette présentation, nous insistons sur l'importance que Warner attachait à la mise en application de ses idées.

¹ Stephen E. Fienberg, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213, U.S.A.; Nuri Jazairi, Department of Economics, York University, North York, Ontario, Canada, M3J 1P3.

5. AUTRES

(1981) Post-treatment randomization in clinical research. *Proceedings of the Social Statistics Section, American Statistical Association*, 233-236.

(1982) Post-treatment randomization extensions for medical and educational research. *Proceedings of the Social Statistics Section, American Statistical Association*, 410-413.

(1983) Post-treatment randomization estimates. *Actes de la 44^{ème} Session, Institut International de Statistique*, Madrid, 464-468.

(1958) Avec R.L. Andriano. Professor Bain and barriers to new competition. *Journal of Industrial Economics*, 66-78.

(1965) Cost models, errors in variables and economies of scale in trucking. *The Cost of Trucking, Econometric Analysts*, Dubuque; William C. Brown and Co., 1-46.

(1983) Avec W.D. Cook et L. Seiford. Preference ranking models: Conditions for equivalence. *Journal of Mathematical Sociology*, 9, 125-137.

Publications et articles principaux de Stanley L. Warner

1928-1992

1. THÈSE DE DOCTORAT ET PUBLICATIONS CONNEXES

- (1962) *Stochastic Choice of Mode in Urban Travel. A Study in Binary Choice*, Evanston, Illinois: Northwestern University Press.
- (1963) Multivariate regression of dummy variates under normality assumptions. *Journal of the American Statistical Association*, 58, 1054-1063.
- (1967) Asymptotic variances for dummy variate regression under normality assumptions. *Journal of the American Statistical Association*, 62, 1305-1314.

- (1965) Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60, 63-69.
- (1971) The linear randomized response model. *Journal of the American Statistical Association*, 66, 884-888.
- (1976) Optimal randomized response models. *Revue Internationale de Statistique*, 44, 205-212.
- (1976) Avec F. Leysieffer. Respondent jeopardy and optimal randomized response models. *Journal of the American Statistical Association*, 71, 649-656.
- (1979) Extended randomized response applications. Dans *Ethical and Legal Problems in Applied Social Research*, (Éds. R. Boruch, J. Ross et J.C. Cecil), Evanston, Illinois: Northwestern University Press.
- (1986) The omitted digit randomized-response model for telephone and other applications. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 441-443.
- (1989) Using randomized response for forecasting dimensions of the AIDS problem. Présentation principale invitée. *The Ninth International Symposium on Forecasting*, Vancouver.
- (1989) Quick randomized response. *Actes de la 47^{ème} Session, Institut International de Statistique*, Paris, Communications libres, 431-432.

3. INFORMATION ÉQUILIBRÉE

- (1975) Advocate scoring for unbiased information. *Journal of the American Statistical Association*, 70, 15-22.
- (1977) Advocate scoring design for technological and social policy assessment. *Actes de la 41^{ème} Session, Institut International de Statistique*, New Delhi, livraison 3 – Communications demandées, 373-379.
- (1979) Subjective information in statistics. *Proceedings of the Business and Economics Section, American Statistical Association*, 558-563.
- (1981) Balanced information, the Pickering Airport experiment. *The Review of Economics and Statistics*, LXII, 256-262.
- (1984) The overlapping information survey model for evaluating summary information. *Proceedings of the Social Statistics Section, American Statistical Association*, 581-584.
- (1985) Applications of the overlapping information model. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 401-403.
- (1985) The overlapping information model for measuring summary information. *Actes de la 45^{ème} Session, Institut International de Statistique*, Amsterdam, Communications libres, 49-50.
- (1987) Identifying rational opinion-formation with the overlapping model. Dans *Applied Probability, Stochastic Processes and Sampling Theory*, (Éds. I.B. MacNeill et G.J. Umphrey), Dordrecht, The Netherlands: D. Reidel, 323-329.
- (1987) Using test populations to develop balanced agenda. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 441-443.
- (1987) *Statistically Balanced Information Technology*. (Manuscrit du livre).

4. ESSAIS CLINIQUES

- (1981) Post-treatment randomization in clinical trials. *Proceedings, Statistics Eighty One Canada*, Concordia University Canada.

d'ordre politique, peut bien souvent empêcher ces dernières de se rencontrer sur un terrain neutre.

L'article de Stephen E. Fienberg et de Nuri Jazairi traite des principales notions autour desquelles s'articulent les travaux de Warner sur le sujet, et que les statisticiens connaissent vraisemblablement moins bien.

Fort répandu aujourd'hui, le système d'annotation musicale élaboré par Stanley Warner, en collaboration avec son épouse, une musicienne, témoigne tout autant de l'esprit inventif de cet homme.

Je ne connaissais pas Stanley Warner à l'époque où il a entrepris ses travaux sur la technique de la réponse randomisée, mais je conserve néanmoins un souvenir précis des échanges que j'ai eus plus tard avec celui-ci au fil de sa carrière. Bien que peu nombreuses, ces conversations m'ont profondément marqué. La chaleur et la modestie de cet homme sans prétentions et fort original ne laissent personne dans l'indifférence. Véritable érudit, Stanley Warner n'aura jamais été un chercheur ordinaire. Convaincu du bien-fondé de ses idées, il n'a pas hésité à emprunter les avenues, parfois bien peu fréquentées, qui s'ouvraient devant lui.

C.-E. Särndal

avec justesse et pertinence des points de vue différents concernant une politique ou une question donnée. Le premier article de Stanley Warner sur le sujet, intitulé "Advocate Scoring for Unbiased Information", a paru dans le *Journal of the American Statistical Association* en 1975. Cet article traite des cas où, à propos d'un sujet ou d'une mesure quelconque, les intervenants dans un débat doivent chacun communiquer des arguments "pour" ou "contre" à un certain nombre d'individus qui, une fois saisis de cette information, doivent ensuite afficher leurs couleurs en tant que partisans du "oui" ou du "non". Chaque intervenant doit préparer ses propres arguments "pour" et "contre" à partir des données disponibles sur le sujet.

Que ce soit au sein des gouvernements ou dans les domaines de l'éducation, du droit, etc., des décisions sont fréquemment prises à la lumière de l'information ainsi communiquée par de tels intervenants. Comme on peut l'imaginer, Stanley Warner était pour sa part préoccupé par le fait que l'information de nature quantitative puisse être utilisée de façon incomplète et parfois arbitraire à l'appui de la prise de décisions de la plus haute importance et ce, même lors de réunions au sommet où le moindre accroc, que ce soit au plan du prestige des parties, de leur méfiance mutuelle, ou pour toute autre considération

Stanley L. Warner

1928-1992

Né aux États-Unis où il fait également ses études, Stanley Warner obtient en 1961 un doctorat en économie de la Northwestern University. En 1971, il s'installe au Canada où il consacra le reste de sa carrière à l'enseignement, en l'occurrence au département d'économie et à la faculté d'administration de la York University. Stanley Warner est décédé soudainement au mois d'août 1992, à l'âge de 63 ans.

Reconnu par ses pairs comme un penseur fort original, les recherches exécutées par Stanley Warner dans le domaine de la statistique ont toujours témoigné de son souci de la pertinence et de son bon sens. En 1994, à Banff, la Société statistique du Canada a tenu un congrès dont une séance a été organisée à sa mémoire. À cette occasion, deux communications (dont on trouvera ci-après un résumé en français) ont été présentées, soit celles de David Bellhouse ainsi que de Stephen E. Fienberg et Nuri Jazairi, respectivement intitulées *Estimation of Correlation in Randomized Response et Stanley Warner's Contributions to Statistically Balanced Information Technology*. Chacune de ces communications portait sur un domaine où Warner a fait oeuvre de pionnier.

Beaucoup, en particulier les statisticiens d'enquête, considèrent Stanley Warner comme l'inventeur de la technique de la réponse randomisée. Cette technique est employée dans le cadre d'enquêtes comportant des questions de nature délicate, par exemple sur la consommation de drogues par les étudiants de niveau secondaire. L'objectif de cette technique est d'éliminer deux erreurs non attribuables à l'échantillonnage et pour le moins gênantes: l'erreur de mesure et la non-réponse. Le biais engendré par ces erreurs n'a cessé de préoccuper les statisticiens depuis l'époque où les premières enquêtes ont été réalisées. Certes, il existe des méthodes normalisées "de correction" du problème, mais celles-ci ne permettent cependant pas d'éliminer les distorsions en question. Faisant fi des idées généralement admises sur la question, Warner s'est attaqué à ces problèmes d'une façon aussi peu conventionnelle que possible pour finalement mettre au point une solution non biaisée et utile, du moins pour certaines enquêtes. Plus tard, en 1965, Warner a publié dans le *Journal of the American Statistical Association* un article qui a fait date, intitulé "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias".

Lorsqu'elle est appliquée dans ce dessein, la technique de la réponse randomisée garantit l'anonymat du répondant. Toute réponse affirmative, formulée par ce dernier, ne saurait permettre son identification, puisque la question visée est sélectionnée de façon aléatoire. Ceci dit, il demeure néanmoins possible (généralement à la faveur d'une augmentation de la variance) d'en arriver à une estimation non

biaisée de la proportion de réponses affirmatives au sein de la population, car le responsable de l'enquête connaît le nombre de réponses affirmatives, de même que la probabilité de sélection au hasard de la question délicate (plutôt qu'une question qui ne le serait pas ou qui serait d'un tout autre ordre).

Comme en fait foi la bibliographie établie en 1988 par Chaudhuri et Mukerjee, qui portait sur les travaux consacrés à la technique de la réponse randomisée, de nombreux statisticiens ont été séduits par les possibilités d'utilisation de cette dernière et lui ont apporté maints perfectionnements et modifications en plus d'en étendre le champ d'application. À quoi tient le vaste intérêt soulevé par cette technique? Il va sans dire que le biais de non-réponse a toujours constitué un important problème d'ordre pratique, auquel on n'avait jusqu'ici trouvé aucune solution satisfaisante, mais là ne réside pas toute l'explication. Il faut également souligner que, même aux yeux des statisticiens d'expérience, cette technique apparaissait purement et simplement magique du fait qu'elle permettait d'obtenir des réponses valides sans même savoir quelle question avait été posée aux répondants. Bien sûr, maintenant que son emploi s'est répandu, il n'est guère difficile d'expliquer pourquoi cette technique a su rallier de nombreux adeptes. De fait, celle-ci est bien souvent citée en exemple aux étudiants, y compris ceux qui suivent des cours d'introduction à la statistique, pour illustrer les forces du raisonnement statistique.

Dans la pratique, l'application de la technique de la réponse randomisée exige le recours à certaines mesures, dont le choix d'une méthode de sélection au hasard d'une question. Au fil des ans, Stanley Warner a compris qu'il devait adapter sa technique aux exigences modernes de collecte de données à faible coût. Toutefois, ce n'est qu'en 1989, dans le cadre de réunions de l'Institut international de statistique tenues à Paris, qu'il a présenté une "technique rapide de réponse randomisée" se prêtant aux sondages téléphoniques, et dans laquelle les répondants répondent aux questions en appuyant sur les touches appropriées du clavier Touch-Tone de leur appareil, ce qui économise du temps et de l'argent.

L'article de David Bellhouse retrace de façon détaillée l'évolution de la technique de la réponse randomisée. Vers 1975, et au cours des années qui ont suivi, Warner s'est principalement intéressé aux méthodes statistiques d'information équilibrée. Au cours des dernières années de sa vie, il travaillait d'ailleurs à la rédaction d'un ouvrage sur le sujet, ouvrage dont le manuscrit en est maintenant rendu à l'étape préparatoire à sa publication.

Le but des techniques d'information équilibrée est de proposer des processus statistiques permettant de présenter

Meeden examine le problème de l'estimation de la médiane lorsqu'une variable auxiliaire est disponible. Il utilise une méthode bayésienne non informative fondée sur une distribution à posteriori de Polya pour les rapports de la variable d'intérêt sur la covariable. L'estimateur ainsi obtenu est comparé d'une manière empirique à un certain nombre de solutions de rechange en termes de biais et d'erreur absolue moyenne pour une gamme de populations réelles et synthétiques. La distribution à posteriori de Polya est également utilisée pour générer des estimations d'intervalles qui font l'objet d'une analyse empirique. L'auteur se penche finalement sur la résistance de la méthode à des écarts modérés par rapport aux hypothèses. Hulliger élabore des estimateurs M basés sur le plan pour des échantillons assortis de probabilités d'inclusion inégales. Il exprime l'estimateur Horvitz-Thompson (HT) sous la forme d'une fonctionnelle des moindres carrés et il en augmente la robustesse contre les valeurs aberrantes par l'intermédiaire d'estimateurs M , selon une méthode analogue à celle de l'augmentation de la robustesse des estimateurs par les moindres carrés dans les modèles linéaires pour des populations infinies. Il présente également une approximation de la variance d'échantillonnage de cet estimateur HT à robustesse accrue ainsi que son estimation. Les résultats de l'étude de Monte-Carlo confirment que les estimateurs HT à robustesse accrue donnent des résultats supérieurs à ceux des estimateurs HT dans nombre de situations où on a affaire à des valeurs aberrantes.

Jachan et Kemp décrivent les plans d'échantillonnage utilisés pour deux enquêtes portant sur les visiteurs dans les parcs relevant du National Park Service au cours d'une période d'un an, et une enquête menée auprès des utilisateurs d'un bassin de trois rivières de la région de Pittsburgh. Ils décrivent les problèmes qui risquent de se poser avec l'échantillonnage dans le temps et dans l'espace, et comparent la façon dont les plans d'échantillonnage de ces deux enquêtes permettent de tenir compte de ces problèmes.

Le rédacteur en chef

Dans ce numéro

Ce numéro de *Techniques d'enquête* comporte une section spéciale à la mémoire de Stanley L. Warner comprenant une introduction présentée par C.-E. Sarnadal, une bibliographie des principales publications et des principaux articles de Warner, classés par sujets, et trois articles abordant des questions où Warner a joué un rôle d'avant-garde. Le premier de ces articles, présenté par Fienberg et Jazairi, résume les travaux de Warner dans le domaine de la technologie de l'information statistiquement équilibrée, c'est-à-dire dans la mise au point de méthodes statistiques propres à fournir une analyse adéquate et équilibrée des divers points de vue exprimés sur une question donnée, à l'occasion d'un débat ou de la prise d'une décision. Les deux autres articles, présentés par Bellhouse et par Mangat et coll., portent sur les réponses randomisées. L'article de Bellhouse débute par un aperçu de la contribution de Warner à l'étude des réponses randomisées et aborde ensuite le problème de l'estimation d'un coefficient de corrélation à partir des données de plans d'échantillonnage à réponses randomisées. Trois plans à réponses randomisées sont examinés: le modèle des questions non liées, celui des constantes additives et celui des constantes multiplicatives. L'article de Mangat et coll. compare l'efficacité des échantillonnages avec ou sans remise dans le contexte des plans à réponses randomisées. Les articles de Fienberg et Jazairi et de Bellhouse s'appuient sur des allocutions présentées lors d'une séance spéciale dédiée à la mémoire de Warner, tenue lors des rencontres de la Société statistique du Canada organisées à Banff, en 1994. Les deux articles suivants, préparés par Lavalée et par Kalton et Brick, portent sur les méthodes de pondération applicables aux estimations transversales effectuées dans le cadre des enquêtes par panel. Lavalée dresse un portrait général de la méthode du partage des poids utilisée pour les estimations transversales des enquêtes longitudinales. Il démontre le caractère non biaisé de cette méthode et obtient une expression générale de la variance de l'estimateur d'un total. Il illustre ensuite cette méthode en l'appliquant au contexte de l'enquête sur la dynamique du travail et du revenu de Statistique Canada. Il traite également de l'estimation de la variance.

Kalton et Brick décrivent des méthodes de pondération utilisées pour les analyses transversales des vagues plus tardives d'une enquête par panel auprès des ménages, lorsqu'on tient compte des données provenant de l'ensemble des ménages ayant fait l'objet de l'enquête. Ces méthodes de pondération peuvent admettre les nouveaux arrivants qui viennent vivre avec les membres de la population originale, mais pas les autres. Les auteurs décrivent les cas où ces méthodes sont optimales et traitent également des ajustements de poids requis pour tenir compte des cas de non-réponse ou de couverture incomplète. Des méthodes d'estimation du sous-dénombrement net des personnes lors de recensements de la population sont abordées par Dick, dans le contexte canadien, et par Kim, Zaslavsky et Blodgett, dans le contexte américain.

L'article de Dick décrit la modélisation réalisée afin de produire des estimations du sous-dénombrement net des personnes à l'intérieur des catégories d'âge, de sexe et de province lors du Recensement canadien de 1991, en utilisant des données provenant de l'étude de contre-vérification des dossiers et de l'étude sur le surdénombrement. Il élabore et utilise un modèle empirique de Bayes pour l'estimation directe des facteurs d'ajustement qui permet un lissage des estimations de ces facteurs d'ajustement. Les estimations du nombre net de personnes manquées sont ensuite transformées à l'aide de la méthode itérative des quotients afin de correspondre aux estimations directes des totaux des groupes nationaux d'âge-sexe et des totaux provinciaux, lesquels sont jugés de bonne qualité.

Kim, Zaslavsky et Blodgett décrivent les deux analyses réalisées pour tester l'hypothèse synthétique de l'homogénéité des taux de sous-dénombrement entre les parties d'états différents qui entrent dans la même post-strate du recensement américain de 1990. Dans la première analyse, la distribution de cinq "variables factices" qui, tout comme le sous-dénombrement, sont liées à la méthode de recensement, sont examinées à l'aide d'un important échantillon tiré du recensement. La deuxième analyse porte sur la distribution du sous-dénombrement et utilise les données du Post Enumeration Survey. Bredt présente des plans fondés sur les chaînes de Markov pour la sélection d'une unité par strate et qui comprennent l'échantillonnage systématique, l'échantillonnage aléatoire simple stratifié et l'échantillonnage systématique équilibré à titre de cas spéciaux. Il présente de nouveaux plans d'échantillonnage qui s'avèrent concurrentiels, sur le plan de l'efficacité de l'estimateur Horvitz-Thompson d'un total, par rapport aux plans classiques à une unité par strate dans toute une gamme de modèles de super-population. Des comparaisons théoriques et numériques sont fournies.

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada
Volume 21, numéro 1, juin 1995

TABLE DES MATIÈRES

Dans ce numéro	1
Hommage à Stanley L. Warner	
Mots d'introduction de C.-E. SÄRNDALE	3
Publications et articles principaux de Stanley L. Warner	5
S.E. FIENBERG et N. JAZAIRI Contributions de Stanley Warner à la technologie de l'information statistiquement équilibrée.....	7
D.R. BELLHOUSE Estimation de la corrélation dans les plans à réponses randomisées	15
N.S. MANGAT, R. SINGH, S. SINGH, D.R. BELLHOUSE et H.B. KASHANI Efficacité de l'emploi des unités non répétées dans une enquête fondée sur la méthode des réponses randomisées	23
<hr/>	
P. LAVALLÉE Pondération transversale des enquêtes longitudinales menées auprès des individus et des ménages à l'aide de la méthode du partage des poids	27
G. KALTON et J.M. BRICK Méthodes de pondération pour les enquêtes par panel auprès des ménages	37
P. DICK Modélisation du sous-dénombrement net dans le recensement du Canada de 1991	51
J.J. KIM, A. ZASLAVSKY et R. BLODGETT Hétérogénéité inter-états du taux de sous-dénombrement et les variables de remplacement dans le recensement des États-Unis de 1990	63
F.J. BREIDT Plans à chaînes de Markov pour l'échantillonnage à une unité par strate.....	73
G. MEBDEN Estimation de la médiane à l'aide d'informations supplémentaires	81
B. HULLIGER Estimateurs Horvitz-Thompson à l'épreuve des valeurs aberrantes	89
R. IACHAN et S.S. KEMP Enquêtes par sondage auprès des visiteurs	99

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

Techniques d'enquête est répertoriée dans The Survey Statistician et Statistical Theory and Methods Abstracts. On peut en trouver les références dans Current Index to Statistics, et Journal Contents in Qualitative Methods.

COMITÉ DE DIRECTION

Président
Membres

- G.J. Brackstone
- D. Binder
- B.N. Chinnappa
- G.J.C. Hole
- F. Mayda (Directeur de la Production)
- M.P. Singh
- C. Patrick
- R. Platak (Ancien président)
- D. Roy

COMITÉ DE RÉDACTION

Rédacteur en chef
Rédacteurs associés

- D.R. Bellhouse, *University of Western Ontario*
- D. Binder, *Statistique Canada*
- M.J. Colledge, *Australian Bureau of Statistics*
- J.-C. Deville, *INSEE*
- J.D. Drew, *Statistique Canada*
- J.-J. Droesbeke, *Université Libre de Bruxelles*
- W.A. Fuller, *Iowa State University*
- M. Gonzalez, *U.S. Office of Management and Budget*
- R.M. Groves, *University of Maryland*
- D. Holt, *University of Southampton*
- G. Kalton, *Westat, Inc.*
- A. Mason, *East-West Center*
- N. Laniel, *M. Latouche, L. Mach, H. Mantel et D. Stukel, Statistique Canada*
- D. Pfeiffermann, *Hebrew University*
- J.N.K. Rao, *Carleton University*
- L.-P. Rivest, *Université Laval*
- I. Sande, *Bell Communications Research, U.S.A.*
- C.-E. Särndal, *Université de Montréal*
- W.L. Schaible, *U.S. Bureau of Labor Statistics*
- F.J. Scheuren, *George Washington University*
- J. Sedransk, *State University of New York*
- P.J. Waite, *U.S. Bureau of the Census*
- J. Waksberg, *Westat, Inc.*
- K.M. Wolter, *National Opinion Research Center*
- A. Zaslavsky, *Harvard University*

POLITIQUE DE RÉDACTION

Techniques d'enquête publie des articles sur les divers aspects des méthodes statistiques qui intéressent un organisme statistique comme, par exemple, les problèmes de conception découlant de contraintes d'ordre pratique, l'utilisation de différentes sources de données et de méthodes de collecte, les erreurs dans les enquêtes, l'évaluation des enquêtes, la recherche sur les méthodes d'enquête, l'analyse des séries chronologiques, la désaisonnalisation, les études démographiques, l'intégration de données statistiques, les méthodes d'estimation et d'analyse de données et le développement de systèmes généralisés. Une importance particulière est accordée à l'élaboration et à l'évaluation de méthodes qui ont été utilisées pour la collecte de données ou appliquées à des données réelles. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de Statistique Canada.

Présentation de textes pour la revue

Techniques d'enquête est publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à faire parvenir le texte rédigé en anglais ou en français au rédacteur en chef, M. M.P. Singh, Division des méthodes d'enquêtes-ménages, Statistique Canada, Tunney's Pasture, Ottawa (Ontario), Canada K1A 0T6. Prière d'envoyer quatre exemplaires dactylographiés selon les directives présentées dans la revue. Ces exemplaires ne seront pas retournés à l'auteur.

Abonnement

Le prix de Techniques d'enquête (n° 12-001 au catalogue) est de 45 \$ par année au Canada, 50 \$ (É.-U.) aux États-Unis, et de 55 \$ (É.-U.) par année à l'étranger. Prière de faire parvenir votre demande d'abonnement à Section des ventes des publications, Statistique Canada, Ottawa (Ontario), Canada K1A 0T6. Un prix réduit est offert aux membres de l'American Statistical Association, l'Association Internationale de Statisticiens d'Enquête et la Société Statistique du Canada.



Ottawa

ISSN 0714-0045

N° 12-001 au catalogue

Autres pays : 55 \$ US

États-Unis : 50 \$ US

Prix : Canada : 45 \$

Juin 1995

Tous droits réservés. Il est interdit de reproduire ou de transmettre le contenu de la présente publication, sous quelque forme ou par quelque moyen que ce soit, enregistré ou non, sans l'autorisation écrite préalable des Services de concession des droits de licence, Division du marketing, Statistique Canada, Ottawa, Ontario, Canada K1A 0T6.

© Ministre de l'Industrie, 1995

Publication autorisée par le ministre
responsable de Statistique Canada

JUN 1995 • VOLUME 21 • NUMÉRO 1

UNE REVUE ÉDITÉE PAR STATISTIQUE CANADA

TECHNIQUES D'ENQUÊTE





NUMÉRO 1

•

VOLUME 21

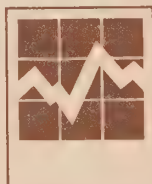
•

JUIN 1995

UNE REVUE
ÉDITÉE
PAR STATISTIQUE CANADA

Catalogue 12-001

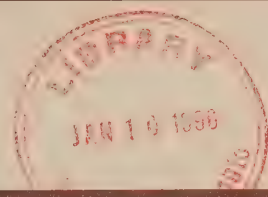
TECHNIQUES D'ENQUÊTE



12
-001



SURVEY METHODOLOGY



Catalogue 12-001

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

DECEMBER 1995

•
VOLUME 21

•
NUMBER 2



Statistics
Canada

Statistique
Canada

Canada



SURVEY METHODOLOGY

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

DECEMBER 1995 • VOLUME 21 • NUMBER 2

Published by authority of the Minister
responsible for Statistics Canada

© Minister of Industry, 1995

All rights reserved. No part of this publication may be reproduced,
stored in a retrieval system or transmitted in any form or by any
means, electronic, mechanical, photocopying, recording or otherwise
without prior written permission from Licence Services,
Marketing Division, Statistics Canada,
Ottawa, Ontario, Canada K1A 0T6.

December 1995

Price: Canada: \$45.00
United States: US\$50.00
Other Countries: US\$55.00

Catalogue No. 12-001

ISSN 0714-0045

Ottawa



Statistics
Canada

Statistique
Canada

Canada

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Survey Methodology is abstracted in The Survey Statistician and Statistical Theory and Methods Abstracts and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods.

MANAGEMENT BOARD

Chairman	G.J. Brackstone	
Members	D. Binder	R. Platek (Past Chairman)
	G.J.C. Hole	D. Roy
	F. Mayda (Production Manager)	M.P. Singh
	C. Patrick	

EDITORIAL BOARD

Editor M.P. Singh, *Statistics Canada*

Associate Editors

D.R. Bellhouse, <i>University of Western Ontario</i>	J.N.K. Rao, <i>Carleton University</i>
D. Binder, <i>Statistics Canada</i>	L.-P. Rivest, <i>Université Laval</i>
J.-C. Deville, <i>INSEE</i>	I. Sande, <i>Bell Communications Research, U.S.A.</i>
J.D. Drew, <i>Statistics Canada</i>	C.-E. Särndal, <i>Université de Montréal</i>
J.-J. Droesbeke, <i>Université Libre de Bruxelles</i>	W.L. Schaible, <i>U.S. Bureau of Labor Statistics</i>
W.A. Fuller, <i>Iowa State University</i>	F.J. Scheuren, <i>George Washington University</i>
M. Gonzalez, <i>U.S. Office of Management and Budget</i>	J. Sedransk, <i>State University of New York</i>
R.M. Groves, <i>University of Maryland</i>	C.J. Skinner, <i>University of Southampton</i>
M.A. Hidiroglou, <i>Statistics Canada</i>	P.J. Waite, <i>U.S. Bureau of the Census</i>
D. Holt, <i>Central Statistical Office, U.K.</i>	J. Waksberg, <i>Westat, Inc.</i>
G. Kalton, <i>Westat, Inc.</i>	K.M. Wolter, <i>National Opinion Research Center</i>
A. Mason, <i>East-West Center</i>	A. Zaslavsky, <i>Harvard University</i>
D. Pfeffermann, <i>Hebrew University</i>	

Assistant Editors J. Denis, M. Latouche, H. Mantel and D. Stukel, *Statistics Canada*

EDITORIAL POLICY

Survey Methodology publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

Survey Methodology is published twice a year. Authors are invited to submit their manuscripts in either English or French to the Editor, Dr. M.P. Singh, Household Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6. Four nonreturnable copies of each manuscript prepared following the guidelines given in the Journal are requested.

Subscription Rates

The price of Survey Methodology (Catalogue No. 12-001) is \$45 per year in Canada, US \$50 in the United States, and US \$55 per year for other countries. Subscription order should be sent to Publication Sales, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6. A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, and the Statistical Society of Canada.

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Volume 21, Number 2, December 1995

CONTENTS

In This Issue	97
P. DAVIS and A. SCOTT The Effect of Interviewer Variance on Domain Comparisons	99
L.-P. RIVEST and D. HURTUBISE On Searls' Winsorized Mean for Skewed Populations	107
L. KISH, M.R. FRANKEL, V. VERMA and N. KAĆIROTI Design Effects for Correlated ($P_i - P_j$)	117
F. DUPONT Alternative Adjustments Where There Are Several Levels of Auxiliary Information ...	125
D.A. BINDER and M.S. KOVACEVIC Estimating Some Measures of Income Inequality from Survey Data: An Application of the Estimating Equations Approach	137
L.R. ERNST and M.M. IKEDA A Reduced-Size Transportation Algorithm for Maximizing the Overlap Between Surveys	147
D.A. DILLMAN, J.R. CLARK and M.D. SINCLAIR How Prenotice Letters, Stamped Return Envelopes and Reminder Postcards Affect Mailback Response Rates for Census Questionnaires	159
L.B. SHRESTHA and S.H. PRESTON Consistency of Census and Vital Registration Data on Older Americans: 1970-1990 ...	167
D. FORSTER and R.W. SNOW An Assessment of the Use of Hand-Held Computers During Demographic Surveys in Developing Countries	179
A.W. SPISAK Statistical Process Control of Sampling Frames	185
Acknowledgements	191

In This Issue

This issue of *Survey Methodology* contains papers on a variety of topics. The first paper, by Davis and Scott, discusses the impact that interviewer effects may have on comparisons between domain means. Using a components of variance model, it is shown theoretically that the impact depends on the distribution of each interviewer's case load between the domains and on the domain-interviewer interaction. The model is applied to data from a health survey to estimate the magnitude of interviewer effects for comparisons between sexes and between ethnic groups. It was found that in some cases the domain-specific interviewer effects have a large impact on the accuracy of between domain comparisons.

Rivest and Hurtubise examine the usefulness of the Winsorized mean as an estimator of the mean of a population that has a distribution skewed to the right. A Winsorized mean is obtained by replacing all observations greater than a given threshold value R by this same value R , before the mean is calculated. The authors suggest a simple algorithm for calculating R that minimizes the squared error of the estimator. They apply this method to several sample sizes and various sample designs, including stratified sampling and sampling with probabilities proportional to size. They derive direct approximations of the effectiveness of the Winsorized mean. They conclude their article with a Monte Carlo simulation to compare various estimators that reduce the impact of extreme values.

Kish, Frankel and Verma examine the possible incidence and the importance of the design effect (deft) on a set of interrelated statistics. On the basis of 14 surveys conducted in six countries, the authors present an empirical approach relating the design effect of analytical statistics, $\text{deft}(p_i - p_j)$, to the design effects of separate statistics, $\text{deft}(p_i)$ and $\text{deft}(p_j)$, for two of the many categories of the same variable. The proposed approximation must be checked constantly. However, it appears to be widely applicable to the data studied, and it is clearly preferable to the hypotheses put forward thus far on $\text{deft}(p_i - p_j)$.

Dupont discusses the estimation of a total from a two-stage sample where auxiliary information is present. First three regression estimators are presented, each making different use of the auxiliary information. Then four calibration estimators are proposed, each corresponding to a specific strategy for using auxiliary information. Dupont then shows that the calibration strategies can be associated with regression modelling. This article also discusses variance estimation for the seven estimators presented, the choice of the estimator where there is nonresponse, and the *a priori* or *a posteriori* use of auxiliary information.

Spisak discusses the use of statistical process control to assure the quality of a frame constructed by a continuous process and used for a survey repeated periodically. The frame sizes constitute a time series for which the appropriate model must be identified in order to estimate the process variance needed for construction of the control charts. The author uses the data from the United States Unemployment Insurance Benefits Quality Control program to illustrate the method.

Binder and Kovacevic show how the estimating equations approach may be used to construct variance estimation procedures that are appropriate when the data come from a survey with a complex design. The approach is most useful when the quantity to be estimated is a complicated non-linear function of the survey population values, as is the case with many common measures of income inequality. Details of the proposed approach are worked out for a number of complex income distribution statistics including the Gini Coefficient, the Lorenz Curve Ordinate, the Quantile Share, and the Low Income Measure. A numerical example is given using data from the Canadian Survey of Consumer Finance.

Ernst and Ikeda present a reduced-size algorithm for maximizing the retention of selected primary sampling units when a new sample (*i.e.*, with a new stratification and allocation) is selected for a repeated survey. First, the transportation procedure developed by Causey, Cox and Ernst (1985) is described. It provides optimal retention of PSU's but the resulting transportation problem may be too large to solve in practice. The authors then expose their algorithm which is an approximation of the previous method but has the advantage of being of smaller size and thus possible to use in many practical situations. Finally, an application of the algorithm to the Survey of Income and Program Participation is presented.

Shrestha and Preston evaluate the consistency of the Census data with the Vital Registration data for the older Americans. First, the data used in the study and their sources of errors are described. Then the authors present the methodology used to evaluate the quality of the old-age statistics and explain how one should interpret the results of the application of that methodology. Finally, results from the application of the methodology to data from 1970 to 1990 are presented.

Dillman, Clark and Sinclair compare different mailout and follow-up strategies with respect to their impact on the response rates for the U.S. Census. The comparison of the strategies includes the use of a factorial design and a sample of 50,000 housing units. The results are analyzed through multiple pairwise comparisons of treatment means and logistic regression.

Forster and Snow evaluate the use of hand-held computers to conduct demographic surveys in developing countries. A data collection test was conducted for comparing the use of paper and computerized questionnaires with the Adult Mortality Survey of people living on the Kenyan coast. The results show that the use of hand-held computers can reduce the data processing time, improve the quality of the data as well as reduce survey costs on the long term.

The Editor

The Effect of Interviewer Variance on Domain Comparisons

PETER DAVIS and ALASTAIR SCOTT¹

ABSTRACT

In this paper we explore the effect of interviewer variability on the precision of estimated contrasts between domain means. In the first part we develop a correlated components of variance model to identify the factors that determine the size of the effect. This has implications for sample design and for interviewer training. In the second part we report on an empirical study using data from a large multi-stage survey on dental health. Gender of respondent and ethnic affiliation are used to establish two sets of domains for the comparisons. Overall interviewer and cluster effects make little difference to the variance of male/female comparisons, but there is noticeable increase in the variance of some contrasts between the two ethnic groupings used in this study. Indeed, the impact of interviewer effects for the ethnic comparison is two or three times higher than it is for gender contrasts. These findings have particular relevance for health surveys where it is common to use a small cadre of highly-trained interviewers.

KEY WORDS: Interviewer variance; Domain comparisons; Design effect.

1. INTRODUCTION

Surveys requiring a high degree of specialist training for interviewers, such as many health studies, are often forced to use a small number of highly-trained interviewers. There has been a substantial amount of work done on estimating the impact of interviewer variability on simple statistics such as means and proportions, and it is well-known that the use of a small number of interviewers, each having a high case load, can lead to a relatively large contribution to the total error. Comprehensive summaries of the literature are given in Groves (1989, chap. 8) and Lessler and Kalsbeek (1992, §11.3). However, most medical and social surveys are primarily interested in more complex questions such as comparisons between sub-groups or estimating the effect of a factor on disease outcome. There is a widespread belief that the effect of interviewer variability is much smaller here, and that the effect of a small number of interviewers is relatively harmless. Following the pioneering work of Kish and Frankel (1974), there has been a great deal of theoretical and empirical work on the effects of clustering on fitting multiple regression models or log-linear models for categorical data. Good accounts of the literature are given in Skinner *et al.* (1989) and Rao and Thomas (1988). There has been some empirical work on the conceptually simpler, yet practically important, problem of comparing sub-group means (see Kish 1987 and Skinner 1989 for example) but relatively little theoretical development.

In this paper we concentrate on comparisons between subgroups (or domains). We first look at theoretical aspects via a straightforward components of variance model. The theory suggests that the impact of interviewer

variability depends on two things, the distribution of each interviewer's case load between the domains and the domain-interviewer interaction. Then we apply the theory to data from a reasonably typical health survey, using two sets of domains defined by the sex and ethnic background of the respondent. Unfortunately the study was not designed *a priori* to estimate interviewer effects (most importantly, interviewers were not deployed at random) so the results should be regarded as suggestive rather than definitive. However, they are sufficiently disturbing to indicate that the problem warrants further study. The results from the ethnic comparisons, in particular, suggest that there are cases when we should be concerned about using a small number of interviewers even when comparisons, rather than simple means or proportions, are the main concern of the analysis.

2. THEORY

For simplicity we start with the special case of a two-stage self-weighting design. This is sufficiently complex to illustrate the central ideas, but simple enough to avoid being swamped with extraneous detail. Following Collins and Butcher (1982), we want to address the problems of interviewer variance and clustering together. A simple correlated response model appropriate for observations drawn according to such a design is

$$Y_{ipr} = \mu + a_i + b_p + e_{ipr}, \quad (1)$$

where i denotes the interviewer, p the primary sampling unit (PSU) and r the individual respondent. Here the

¹ Peter Davis, Department of Community Health, University of Auckland, Private Bag 92019, Auckland, New Zealand; and Professor Alastair Scott, Department of Mathematics and Statistics, University of Auckland, Private Bag 92019, Auckland, New Zealand.

mean, μ , is fixed constant and the remaining components, a_i , b_p and e_{ipr} , are assumed to be independent random variables with variances σ_I^2 , σ_C^2 and σ^2 respectively. Such models have been used widely in theoretical studies of response variance. See Prasad and Rao (1990) for a recent example. For references to earlier work, see the comprehensive treatment in §11.3 of Lessler and Kalsbeek (1992).

Since the design is self-weighting the sample mean, \bar{Y} , is the natural estimator of the population mean. Its variance under the correlated response model (1) is

$$V(\bar{Y}) = (\bar{n}_I \sigma_I^2 + \bar{n}_C \sigma_C^2 + \sigma^2)/n$$

with $\bar{n}_I = \sum_i n_i^2/n$, where n_i is the number of respondents handled by the i -th interviewer and $n = \sum_i n_i$ is the total sample size, and $\bar{n}_C = \sum_p m_p^2/n$ where m_p denotes the number of respondents in the p -th PSU. Note that \bar{n}_I is always larger than the simple arithmetic average of the n_i 's and can be considerably larger if the n_i 's vary widely.

Now consider what the corresponding expected variance, $V_0(\bar{Y})$ say, would be if the n observations had been generated independently (e.g. if we had drawn a simple random sample from a very large population of PSUs using a large pool of interviewers). It follows from (1) that

$$V_0 = \sigma_{i0t}^2/n \quad (2)$$

where

$$\sigma_{i0t}^2 = \sigma_I^2 + \sigma_C^2 + \sigma^2.$$

The inflation in the expected variance due to the combined effects of interviewer variability and intra-cluster correlation is given by the ratio

$$\begin{aligned} D_0 &= V(\bar{Y})/V_0 \\ &= 1 + (\bar{n}_I - 1)\rho_I + (\bar{n}_C - 1)\rho_C \end{aligned} \quad (3)$$

where $\rho_I = \sigma_I^2/\sigma_{i0t}^2$ and $\rho_C = \sigma_C^2/\sigma_{i0t}^2$. We shall refer to this ratio as the “design effect” although it differs slightly from the usual definition which is in terms of actual, rather than expected, variances. It is clear from expression (3) that interviewer variability can have a substantial effect on the variance of a sample mean if the average interviewer case-load, \bar{n}_I , is large even if the intra-interviewer correlation, ρ_I , is relatively small.

Next suppose that we are interested in the difference between two domain means rather than a single mean. We might, for example, be interested in gender differences or in differences between two ethnic groups. In the simplest extension of the correlated response model (1) we might postulate a model of the form

$$Y_{ipr}^{(d)} = \mu^{(d)} + a_i + b_p + e_{ipr}^{(d)} \quad (4)$$

for observations from the d -th domain. Here the means, $\mu^{(d)}$, may be different for the two domains but the interviewer and cluster effects are assumed to be the same.

Let $p_i^{(d)} = n_i^{(d)}/n^{(d)}$, where $n_i^{(d)}$ is the number of respondents from domain d contacted by the i -th interviewer and $n^{(d)}$ is the total number of respondents from domain d . Similarly, let $q_p^{(d)} = m_p^{(d)}/n^{(d)}$, where $m_p^{(d)}$ is the number of respondents from domain d lying in the p -th PSU. Then, under model (4), the expected variance of $\bar{Y}^{(a)} - \bar{Y}^{(b)}$, the difference between the sample means for the two domains, is

$$V(\bar{Y}^{(a)} - \bar{Y}^{(b)}) = (\bar{m}_I \sigma_I^2 + \bar{m}_C \sigma_C^2 + \sigma^2) \left(\frac{1}{n^{(a)}} + \frac{1}{n^{(b)}} \right), \quad (5)$$

where

$$\bar{m}_I = \sum_i (p_i^{(a)} - p_i^{(b)})^2 / \left(\frac{1}{n^{(a)}} + \frac{1}{n^{(b)}} \right) \quad (6)$$

and

$$\bar{m}_C = \sum_p (q_p^{(a)} - q_p^{(b)})^2 / \left(\frac{1}{n^{(a)}} + \frac{1}{n^{(b)}} \right). \quad (7)$$

If the observations had been generated independently the corresponding expected variance would be

$$V_1 = \sigma_{i0t}^2 \left(\frac{1}{n^{(a)}} + \frac{1}{n^{(b)}} \right)$$

so that the inflation due to interviewer variability and intra-cluster correlation is now

$$\begin{aligned} D_1 &= \text{Var}(\bar{Y}^{(a)} - \bar{Y}^{(b)})/V_1 \\ &= 1 + (\bar{m}_I - 1)\rho_I + (\bar{m}_C - 1)\rho_C. \end{aligned} \quad (8)$$

The size of the effect depends on the way the interviewers' case-loads and the PSUs cut across the domains. At one extreme, when each interviewer contacts the same proportion of people from both domains, (i.e. when $p_i^{(a)} = p_i^{(b)}$), \bar{m}_I is zero and the interviewer effect essentially cancels out. At the other extreme, when each interviewer sees only cases from a single domain, \bar{m}_I is similar in size to \bar{n}_I and the interviewer effect for differences is comparable to that for a single mean. Typically interviewers contact people from both domains and \bar{m}_I is rather small, giving some justification to the belief that interviewer variability has a small impact on estimated differences between domains. Similar comments apply to the effect of clustering.

All this depends on the assumption that the interviewer and cluster effects, a_i and b_p , are the same for both domains. It is easy to imagine situations where such an assumption would not be at all reasonable. Some interviewers, for example, might interact very differently with males and females, or with members of different ethnic groups. A model which allows for the possibility of such interactions is

$$Y_{ipr}^{(d)} = \mu^{(d)} + a_i^{(d)} + b_p^{(d)} + e_{ipr}^{(d)}, \quad (9)$$

where $a_i^{(a)}$ and $a_i^{(b)}$ (respectively $b_p^{(a)}$ and $b_p^{(b)}$) are now assumed to be correlated random variables with correlation $r_I(r_C)$. The naive model (4) corresponds to the special case in which the variances of the effects are equal and r_I and r_C are both equal to one. On the other hand, if there are substantial differences between the interviewer (cluster) effects for the two domains, $r_I(r_C)$ will be small (or even negative in extreme cases). In the rest of this section we suppose for simplicity that the variances of $a_i^{(a)}$ and $a_i^{(b)}$ (respectively $b_p^{(a)}$ and $b_p^{(b)}$) are equal. This may or may not be reasonable in practice but the simplification enables us to concentrate on the essential ideas. The basic form is similar in the more general case but the terms are somewhat messier. Under model (9), the expected variance of $\bar{Y}^{(a)} - \bar{Y}^{(b)}$ is

$$V(\bar{Y}^{(a)} - \bar{Y}^{(b)}) = (\bar{v}_I \sigma_I^2 + \bar{v}_C \sigma_C^2 + \sigma^2) \left(\frac{1}{n^{(a)}} + \frac{1}{n^{(b)}} \right) \quad (10)$$

where

$$\bar{v}_I = \sum_i (p_i^{(a)2} - 2r_I p_i^{(a)} p_i^{(b)} + p_i^{(b)2}) \left/ \left(\frac{1}{n^{(a)}} + \frac{1}{n^{(b)}} \right) \right.$$

with a similar definition for \bar{v}_C in terms of $q_p^{(a)}$ and $q_p^{(b)}$.

The variance inflation factor under this model is

$$D_2 = 1 + (\bar{v}_I - 1)\rho_I + (\bar{v}_C - 1)\rho_C. \quad (11)$$

This is a decreasing function of r_I , the correlation between the interviewer effects for the two domains; the smaller the correlation, the larger the variance inflation. When $r_I = 1$, \bar{v}_I reduces to \bar{m}_I and the interviewer effect is negligible provided all interviewers see a reasonable balance of people from both domains. However, if r_I is small (indicating a strong interaction between the interviewers and domains), \bar{v}_I is the same order of magnitude as \bar{n}_I and the effect of interviewer variability on the variance of domain differences can be substantial.

In practice, the effects will fall between the two extremes and their likely impact is a matter for empirical enquiry. In the next section, therefore, we make a start on building up practical knowledge about the impact using data for

a variety of questions drawn from a single health survey that is typical of the genre of research investigation for which domain comparisons are important (although not ideally designed for our purposes!).

3. EXAMPLE

The example is based on data drawn from a survey of the oral health, attitudes and practices of adult New Zealanders. The details of the survey are reported in full elsewhere (Cutress *et al.* 1979). The important features of the study for the purposes of the current investigation are the sample design and the deployment of interviewers.

The sample design was a stratified multi-stage sampling scheme. The country was divided into 256 Territorial Local Authorities (TLAs) and a geographically stratified sample of 68 TLAs was drawn from the 256 with selection probabilities proportional to size (PPS) at the first stage, where size was the estimated number of persons aged 15 and over. Each sampled TLA was split into secondary sampling units (SSUs) comprising existing census mesh-blocks, aggregated where necessary in order to achieve a minimum size of 50. Two SSUs were then selected with PPS from each sampled TLA at the second stage. Finally, a systematic sample of 28 adults was drawn from each sampled SSU. This equalised the final probability of selection for all adults so that the sample design is (approximately) self-weighting.

The key point of the design was the deployment of the interviewers. Thirteen interviewers were employed in the study, with at least three interviewers used within each SSU, and all interviewers carried out at least 10% of their total work-load in one region (Auckland). Ideally the assignment of interviewers would be part of the overall sample design as in Fellegi (1974) or Biemer and Stokes (1985). Unfortunately the study was not designed to estimate interviewer variance, and the assignment of respondents to interviewers was done in a haphazard way, rather than using a formal randomization procedure.

This is fairly typical of large studies. The following quote from Hox (1994) gives a good summary of the situation: "Ideally, in interviewer studies, respondents should be assigned to interviewers at random. In large-scale studies, this is seldom done because it is expensive and complicated to organize. This makes it difficult to use such studies for methodological research because, as a result, interviewer and respondent characteristics might be confounded. Multi-level analysis, as outlined above, offers some remedies for this situation. If the relevant respondent variables are known they can be put in the regression model to equalize interviewers by statistical means. . . The limitation of this approach is that it relies on statistical control instead of experimental control. It depends on the assumption that all relevant covariates have been included

and have been correctly modeled. Without randomization, it is impossible to conclude that the influence of all confounding variables has been eliminated.” In our case, the deployment is such that all the components of variance are formally identifiable, provided that we believe the model and are willing to accept that the assignment of interviewers is independent of the cluster effects. However, because of the lack of formal randomization there always remains the possibility that variations in patterns of response between interviewers could be a function of workload allocation rather than interviewing style. Clearly the empirical results can only be regarded as tentative, pointing out possibilities that will need to be explored further in properly designed studies.

Even if we ignore the lack of randomization in the interviewer deployment, the study design is considerably more complicated than the one assumed for the development of the theory in the previous section, since it involves three stages of sampling and regional stratification of the first stage units. In the full analysis, we fitted a more complex model including fixed effects terms for the stratification, a hierarchical random effects model for the three stages of sampling, and all second-order interaction terms. However it turned out that the TLAs used as the first stage units were so diffuse that the differences between strata and the between-TLA component of variance were negligible for all the variables used in the following analysis. Thus the between-SSU component is dominant and, for all practical purposes, we can treat the design as if it were a two-stage sample with the meshblocks (aggregated where appropriate) as PSUs. We have ignored the other components in the results reported in the next section.

4. RESULTS

We look first at interviewer and cluster effects on a selection of means and proportions. We have used Model (9) for both types of variable. It is now well-known that this leads to an under-estimate of the variance components for binary data (see Anderson and Aitken 1985 and Pannekock 1988 for example), so our estimated design effects for proportions should be regarded as lower bounds. The models are fitted using PROC GLM in SAS. The impact of clustering has been well documented in the literature (Kish 1965; Kish and Frankel 1970; 1974). In general terms, the magnitude of the effects of clustering depends on the type and number of units chosen and is likely to vary with different kinds of social and demographic characteristics. In the current investigation clustering effects were expected to be reasonably high because the census meshblocks used as sampling units are likely to show a fair degree of internal homogeneity. In keeping with this concentration of population characteristics, it was assumed that demographic and related items would show the largest values of ρ_C . Values of ρ_I were expected

to be lower because of the intense interviewer training. The literature suggests that these effects are also likely to vary according to the type of questionnaire item, with attitude questions, questions requiring probing, fixed-alternative and forced-choice items, together with poorly-worded and ambiguous questions, being particularly susceptible to interviewer variability (Feather 1973, Groves 1989).

Estimated measures of intra-interviewer and intra-cluster correlation coefficients for a selection of questionnaire items falling under four separate headings (socio-demographic, attitudinal, reports of recent behaviour, and recall of distant behaviour) are outlined in the first two columns of Table 1. These categories were identified as providing natural groupings with the potential to display a wide range of interviewer effects. Within each grouping the items are listed in order of the size of their intra-interviewer correlations. A full description of all questionnaire items (apart from the self-evident socio-demographic category) is provided in the Appendix.

As expected, the socio-demographic variables (except for gender) show the highest values of intra-cluster correlation. The average ρ_C is .07 (.08 if gender is omitted). The average values of ρ_C for the other three categories of item are .02 and less. A few items that might be expected to be closely related to social background – like dental visiting, payment for visits, toothbrushing and certain attitude statements – have higher than average ρ_C values. In general, though, these values fall within the range reported by others. (See, for example, Kish 1965, p. 581 for a series of consumer surveys, Bebbington and Smith 1977 and Verma *et al.* 1984 for the country studies in the World Fertility Survey.)

The corresponding estimated ρ_I values are listed in the first column of Table 1. In general these values are very much smaller than those recorded for cluster effects, being usually less than half, and in some cases a tenth, the size of the ρ_C values for the corresponding items. As expected, some attitude items show higher than average ρ_I values, as do certain reports of behaviour that might be susceptible to a high “social desirability” bias, like toothbrushing and buying sweets and chocolates. Ethnic group and employment status also record relatively high values. The pattern is similar to that found in previous studies, although the values recorded lie at the lower end of the range of typical values reported elsewhere (Feather 1973; Kish 1962; O’Muircheartaigh 1977; O’Muircheartaigh and Wiggins 1981). A comprehensive survey is given in Chapter 8 of Groves (1989). This may partly reflect the intensive training and monitoring of the interviewers that were integral to the field work stage of the study. It may also be influenced by the rigorous post-field work “cleaning” (editing and checking) of the data that was carried out prior to analysis. However it may also simply be due to the attenuation resulting from using Model (1) for proportions that we noted above.

Table 1

Cluster and Interviewer Effects for Means and Proportions

Item Description	$\hat{\rho}_I$	$\hat{\rho}_C$	D_0	%Int
Attitudinal:				
Dentists 1	.014	.014	4.61	91
Visiting	.008	.028	3.42	74
Natural Teeth	.008	.027	3.52	76
Health of Teeth	.007	.015	2.97	84
Dentures	.005	.015	2.67	80
Dentists 2	.004	.033	2.77	57
Health of Gums	.003	.010	1.96	77
Fluoridation	.001	.016	1.66	49
Average	.006	.020	2.95	73
Socio-demographic:				
Employment Status	.010	.055	4.20	65
Race	.009	.172	6.87	33
Age	.004	.042	5.98	52
Household Income	.002	.092	3.29	15
Marital Status	.000	.058	2.34	0
Sex of Respondent	.000	.005	1.12	0
Average	.004	.071	3.47	28
Recent Behaviour:				
Brushed Teeth	.019	.025	6.16	8
Sweets/Chocolates	.011	.003	3.75	98
Fluoride Toothpaste	.008	.000	3.04	100
Toothpick	.006	.006	2.66	92
Rinse Mouth	.004	.024	2.43	62
Dental Floss	.001	.018	1.60	43
Disclosing Tablet	.000	.027	1.49	0
Mouthwash	.000	.018	1.42	0
Average	.006	.012	2.82	60
Distant Behaviour:				
Age First Paid	.004	.029	2.34	57
Visited Dentist	.004	.029	2.51	57
Cost Last Year	.002	.000	1.19	100
Year Last Visit	.000	.014	1.15	0
Average	.003	.018	1.80	54

Perhaps more significant than the pattern and values of ρ_I is the impact of interviewer variability on the overall design effect, incorporating both interviewer and clustering effects. This is shown in the third column of Table 1 (D_0), with the final column (% Int) representing the proportionate contribution of interviewer variability to the overall value of D_0 . Design effects are substantial, being above two in all but a minority of cases. This is due to the clustering and to the impact of the large interviewer workloads characteristic of the study since, from equation (3), the variance is increased by a factor of $1 + (\bar{n}_I - 1)\rho_I$, where \bar{n}_I is a weighted average of the interviewer workloads. There is a distinct pattern in the contribution to the design effect produced by interviewer variability. For socio-demographic variables it averages just under one half of the contribution from clustering, while for attitudinal

items the interviewer contribution to the design effect rises to three times that from clustering. The other two categories of items range in between these two extremes.

What the results outlined in Table 1 confirm is the impact that interviewer workload has on the variance of sample estimates, because of the multiplier effect. In essence, an interviewer component with a very small intra-class correlation can be translated into a major effect if the interviewer workload is high. In the study under review, the logistics of deployment and the requirements of on-going quality control seemed to argue for small interview teams, a practice that appears to be typical of much field work in the health area (for example, Choi and Comstock 1975). This meant that interview workloads averaged over 250. The cost of this strategy is immediately apparent from the results in Table 1; very small differences between interviewers are translated into major reductions in the precision of sample estimates.

Now we turn to the main object of our analysis, viz. the impact of interviewer variability on contrasts between domain means or proportions. In the current analysis, this was assessed for two sets of comparisons, the first set by gender (male/female) and the second set by ethnic group (European/non-European). As we have seen in the discussion following equation (11), the contribution, $1 + (v_I - 1)\rho_I$, to D_0 from interviewer differences depends on the extent to which the interviewer effect is constant across the two domains and on the way the domains cut across individual case-loads. Assuming that the domains cut evenly across interviewer case-loads, then v_I is zero if the interviewer effect is identical in the two domains, in which case the common interviewer effect cancels out completely in the comparison. On the other hand, if the effects in the two domains are weakly correlated then the value of v_I tends to be much higher and in extreme cases may equal the average case-load. In the current study values of v_I fell between 0 and 50 for both gender and ethnic group. Thus the effect of domain-specific interviewer effects on the design effect can be quite substantial. Similar comments apply to the impact of clustering on the comparison; if the effect is the same on both domains then it largely cancels out and the net impact is small, but the impact can be substantial if the clustering effect is domain specific.

Table 2 shows values of ρ_I and ρ_C for comparisons by gender and by ethnic group, together with the overall design effect D_2 and the proportion of this effect due to the impact of interviewer variability. Note that the item on the use of disclosing tablets has been omitted from Table 2. This is because so few respondents either used or knew what this item was that the effective sample size in this case is tiny, thus rendering the results almost meaningless.

The impact of both interviewer and clustering on comparisons by gender is small with design effects little above unity, in spite of the fact that the estimated values of ρ_I and ρ_C are slightly increased when adjusted for this variable.

Table 2
Interviewer and Cluster Effects for
Domain Differences

Item Description	By Sex				By Race			
	$\hat{\rho}_I$	$\hat{\rho}_C$	D_2	%Int	$\hat{\rho}_I$	$\hat{\rho}_C$	D_2	%Int
Attitudinal:								
Dentists 2	.004	.043	1.05	0	.010	.027	1.28	46
Visiting	.009	.028	1.08	0	.072	.133	5.19	78
Natural Teeth	.010	.032	1.06	0	.010	.037	1.26	42
Fluoridation	.001	.019	1.12	42	.021	.031	2.13	88
Dentures	.007	.018	1.04	0	.011	.035	1.21	33
Health of Teeth	.012	.022	1.52	85	.010	.045	1.40	20
Dentists 1	.001	.018	1.05	0	.015	.020	1.53	74
Health of Gums	.006	.022	1.07	0	.003	.104	1.46	9
Average	.006	.025	1.12	16	.019	.054	1.93	49
Socio-demographic:								
Race	.008	.183	1.11	0	–	–	–	–
Household Income	.004	.095	1.37	24	.004	.099	1.95	40
Marital Status	.000	.059	1.17	0	.011	.060	1.69	38
Employment Status	.014	.067	1.42	71	.022	.116	2.09	25
Age	.007	.052	1.06	0	.006	.093	1.87	24
Sex of Respondent	–	–	–	–	.006	.011	1.09	44
Average	.007	.091	1.23	19	.010	.076	1.74	34
Recent Behaviour:								
Brushed Teeth	.025	.060	1.62	65	.019	.019	1.68	88
Rinse Mouth	.007	.029	1.28	64	.004	.023	1.20	45
Mouthwash	.000	.057	1.20	0	.027	.105	2.69	75
Dental Floss	.003	.021	1.06	0	.015	.036	1.37	32
Toothpick	.006	.010	1.03	0	.006	.046	1.48	63
Sweets/Chocolates	.012	.009	1.02	0	.013	.022	1.31	48
Fluoride Toothpaste	.010	.007	1.11	100	.007	.000	1.02	100
Average	.009	.028	1.19	33	.013	.036	1.36	64
Distant Behaviour:								
Age First Paid	.003	.033	1.10	0	.029	.141	2.92	71
Visited Dentist	.005	.035	1.04	0	.020	.018	1.26	50
Year Last Visit	.004	.012	1.20	75	.016	.003	1.83	12
Cost Last Year	.007	.021	1.01	0	.076	.117	2.09	42
Average	.005	.025	1.09	19	.035	.070	2.03	44

A significant gender-specific effect was apparent for only three items, health of teeth and tooth-brushing – for which there may be a unique social acceptability bias – and employment status – which holds quite different connotations for men and women. Note that the interviewer effect is the dominant one in all three of these comparisons.

The impact on comparisons by ethnic group is much higher, with design effects averaging about 1.7. This suggests that there are significant, non-cancelling interviewer and clustering effects associated with the ethnic identity of respondents. There are large ethnic-specific interviewer effects for two hypothetical attitudinal questions (visiting and fluoridation), for one item of recent behaviour, and for age of first payment for dental services. The result is plausible; all the interviewers were European and may have varied systematically in their interactions with respondents of different ethnic backgrounds. Again clustering effects are most marked for the socio-demographic variables. Not only are the design effects on average higher than those recorded for the gender comparisons, but the interviewer component is in general two or three times higher for the ethnic group contrasts.

A referee rightly points out that because of the way the interviewers are deployed (they worked primarily in teams assigned to different parts of New Zealand), there is a real possibility that the interviewer effects might be inflated because of confounding with area effects. The fact that differences between the TLAs were so small gives us some reason to believe that this inflation will be small, but the possibility can never be discounted with this design.

5. DISCUSSION

This paper has applied empirical data from a not untypical health survey to assess the impact of interviewer variability under the assumptions of both simple and extended versions of the correlated response model for the error variance of a multi-stage sample design.

In the first case the simple model analyses the relative impact of cluster and interviewer effects on the estimation of means and proportions. The results of this analysis confirm a number of findings that are well established in the literature: the intra-class correlations for interviewers are generally lower than those for clusters; the intra-class correlations for clusters vary in the expected direction by question type; the overall design effects for these question types vary between 2 and 3.5; a substantial component of this inflation is contributed by interviewer variability and can probably be attributed to the multiplier effect of large interviewer caseloads; finally, the impact of this interviewer component is shown to vary in the expected direction by question type.

In the second case the extended model addresses the analysis of cluster and interviewer effects for the estimation of domain contrasts between means and proportions for two sets of comparisons defined by gender and ethnic group. The effect on contrasts between domain means was smaller but it was still significant for a number of items, particularly for the ethnic comparisons, suggesting that the interviewer effect was different for the two domains. The size of the effect for these items was certainly large enough to suggest that we should be concerned about it in designing similar studies. In general, the impact of interviewer effects was two or three times as great for the ethnic contrasts as it was for the gender comparisons.

The basic deficiencies in the design mean that these results must be regarded as suggestive rather than definitive. They do indicate, however, that there is considerable potential for damage in the use of a small group of interviewers even when interest is centered on domain differences rather than simple means or proportions. This is certainly counter to standard folklore in some fields such as health surveys, and suggests that considerable further empirical work is justified.

On the assumptions of the simple correlated response model a reduction in the impact of interviewer variance

can be achieved by raising the number of interviewers and thus reducing individual interviewer workloads. Of course, this brings with it a potential reduction in the quality of interviewing if training and monitoring procedures have to be tempered. In this instance close attention to question wording and interviewer instruction is clearly crucial. In the case of the extended version of the correlated response model, however, such a strategy is unlikely to be a sufficient one on its own. If comparisons between groups are a major objective of the study, then it is important also to ensure that the interviewers treat the two groups in as similar a way as possible. It is also important to design the study so that each interviewer contacts respondents drawn from both groups. This is likely to be a critical consideration in investigations such as case-control studies in which health outcomes are related to contrasting exposures and in which the control of potential confounder variables may have a significant influence on the magnitude of measures such as the odds ratio.

ACKNOWLEDGEMENTS

This project has received support from the Medical Research Council of New Zealand. The bulk of the detailed computations for this paper were carried out by Joanna Broad.

APPENDIX

Questionnaire Items

Attitudinal

- Dentists 1: "Dentists are more interested in their patients than making money."
- Dentists 2: "Dentists recommend a lot more things to be done than really need to be done."
- Dentures: "Dentures are just as good (or better) than your own teeth."
- Fluoridation: "What is your opinion on fluoridating public water supplies?"
- Visiting: "Do you think a person should go to the dentist only when they have dental problems or should they go sometimes also when they have no obvious problems?"
- Health of Teeth: "If you went to the dentist tomorrow, do you think he would find anything wrong with your teeth?"
- Health of Gums: "If you went to the dentist tomorrow do you think he would find anything wrong with your gums?"

Recent Behaviour

- "Yesterday did you – use a disclosing tablet/mouthwash/dental floss/toothpick?
– rinse after eating?
– brush your teeth?"

"Did you buy sweets or chocolates any time last week?"

Distant Behaviour

- Age First Paid: "About how old were you when you first went to a dentist for routine treatment for which you or your family had to pay?"
- Visited Dentist: "Did you visit a dentist in the last 12 months?"
- Year Last Visit: "In what year did you last visit a dentist?"
- Cost Last Year: "About how much did you pay for dental treatment in the last 12 months?"

REFERENCES

- ANDERSON, D.A. (1985). Variance component models with binary response; interviewer variability. *Journal of the Royal Statistical Society, Series B*, 47, 203-210.
- BIEMER, P.B., and STOKES, S.L. (1985). Optimal design of interviewer variance experiments in complex surveys. *Journal of the American Statistical Association*, 80, 158-166.
- BEBBINGTON, A.C., and SMITH, T.M.F. (1977). The effect of survey design on multivariate analysis, *The Analysis of Survey Data*, (C.A. O'Muircheartaigh and C. Payne, Eds.), 175-192. New York: John Wiley.
- CHOI, I.C., and COMSTOCK, G.W. (1975). Interviewer effects on responses to a questionnaire relating to mood. *American Journal of Epidemiology*, 101, 84-92.
- COLLINS, M., and BUTCHER, B. (1992). Interviewer and clustering effects in an attitude survey. *Journal of the Market Research Society*, 25, 39-58.
- CUTRESS, T.W., HUNTER, P.B., DAVIS, P.B., BECK, D.J., and CROXSON, L.J. (1979). *Adult Oral Health and Attitudes to Dentistry in New Zealand*, Medical Research Council, Wellington.
- DIJKSTRA, W. (Ed.) (1982). *Response Behaviour in the Survey Interview*. New York: Academic Press.
- FEATHER, J. (1973). *A Study of Interviewer Variance*, (WHO/ICS-MCU Saskatchewan Study Area Reports Series 2, No. 3). Department of Social and Preventive Medicine. University of Saskatchewan, Saskatoon.
- FELLEGI, I.P. (1974). An improved method of estimating correlated response variance. *Journal of the American Statistical Association*, 69, 496-501.
- GROVES, R. (1989). *Survey Errors and Survey Costs*. New York: John Wiley.

- GROVES, R., and FULTZ, N.H. (1985). Gender effects among telephone interviewers in a survey of economic attitudes. *Sociological Methods and Research*, 14, 31-52.
- HOLT, D., SMITH, T.M.F., and WINTER, P.D. (1980). Regression analysis of data from complex surveys. *Journal of the Royal Statistical Society*, 143, 474-487.
- HOX, J.J. (1994). Hierarchical regression models for interviewer and respondent effects. *Sociological Methods and Research*, 22, 300-318.
- KISH, L. (1962). Studies of interviewer variance for attitudinal variables. *Journal of the American Statistical Society*, 57, 92-115.
- KISH, L. (1965). *Survey Sampling*. New York: John Wiley.
- KISH, L., and FRANKEL, M.R. (1974). Inferences from complex samples. *Journal of the Royal Statistical Society, Series B*, 36, 1-37.
- KISH, L. (1987). *Statistical Design for Research*. New York: John Wiley.
- LESSLER, J.T., and KALSBEEK, W.D. (1992). *Nonsampling Errors in Surveys*. New York: John Wiley.
- O'MUIRCHEARTAIGH, C.A. (1976). Response errors in an attitudinal sample survey. *Quality and Quantity*, 10, 97-115.
- O'MUIRCHEARTAIGH, C.A., and PAYNE, C. (Eds.) (1977). *The Analysis of Survey Data*, (Volume 2: Model Fitting). New York: John Wiley.
- O'MUIRCHEARTAIGH, C.A., and WIGGINS, R.D. (1981). The impact of interviewer variability in an epidemiological survey. *Psychological Medicine*, 11, 817-824.
- PANNEKOEK, J. (1988). Interviewer variance in a telephone survey. *Journal of Official Statistics*, 4, 375-384.
- PRASAD, N.G., and RAO, J.N.K. (1990). The estimation of mean squared error of small area estimators. *Journal of the American Statistical Association*, 85, 163-171.
- RAO, J.N.K., and THOMAS, D.R. (1988). The analysis of cross-classified data from complex sample surveys. *Sociological Methodology*, 18, 213-269.
- SKINNER, C.J., HOLT, D., and SMITH, T.M.F. (Eds.) (1989). *Analysis of Complex Surveys*. New York: John Wiley.
- VERMA, V., SCOTT, C., and O'MUIRCHEARTAIGH, C.A. (1980). Sample designs and sampling errors for the World Fertility Survey. *Journal of the Royal Statistical Society, Series A*, 143, 431-473.

On Searls' Winsorized Mean for Skewed Populations

LOUIS-PAUL RIVEST and DANIEL HURTUBISE¹

ABSTRACT

This paper considers the winsorized mean as an estimator of the mean of a positive skewed population. A winsorized mean is obtained by replacing all the observations larger than some cut-off value R by R before averaging. The optimal cut-off value, as defined by Searls (1966), minimizes the mean square error of the winsorized estimator. Techniques are proposed for the evaluation of this optimal cut-off in several sampling designs including simple random sampling, stratified sampling and sampling with probability proportional to size. For most skewed distributions, the optimal winsorization strategy is shown, on average, to modify the value of about one data point in the sample. Closed form approximations to the efficiency of Searls' winsorized mean are derived using the theory of extreme order statistics. Various estimators reducing the impact of large data values are compared in a Monte Carlo experiment.

KEY WORDS: Outliers; Max domain of attraction; Mean square error; Simple random sampling; Stratified sampling.

1. INTRODUCTION

Samples drawn from positively skewed populations often contain outliers with values that are much larger than most sampled values. One usually tries to accommodate these large values when designing the survey (Glasser 1962; Hidioglou 1987). However, given the multipurpose nature of most surveys, statisticians are often faced with outliers at the estimation stage. These data points make classical survey estimators, such as the sample mean, unstable. It is therefore of interest to study alternative estimators that lower the impact of large data values. Winsorization (Searls 1966) consists in replacing the data values larger than a cut-off value R by R before averaging. Searls suggested to select the value of R which minimizes the mean square error of the winsorized mean. One can also take R equal to the second largest data value in the sample (Rivest 1994). Searls' estimator was best among all the methods to adjust large data values studied by Ernst (1980). Hicks and Fetter (1993) implement Searls' winsorization strategy in an agriculture survey. Other strategies have been proposed for dealing with large observations in survey sampling. Chambers and Kokic (1993) review estimators derived from the theory of "Robust Statistics" (Huber 1981). Fuller (1991, 1993) proposes a preliminary test to detect the presence of extreme values in the sample; the impact of these values is lowered only in samples for which this test is significant. Lee (1994) provides a good review of this expanding literature.

The key to the implementation of Searls' winsorization method is the selection of the cut-off R . A simple algorithm for calculating the optimal cut-off for a known population

in simple random sampling and in pps sample is proposed in Section 2. Repeated calculations of the optimal cut-off for several populations and several sample sizes reveal that, in most cases, the optimal scheme winsorizes one data point on average, regardless of the sample size. Section 3 extends the result of Section 2 to stratified sampling. A simple algorithm for the calculation of cut-off values in each stratum is proposed. The rule of winsorizing an average of one data point per sample regardless of sample size is shown to hold also in stratified samples. The efficiencies, with respect to the sample mean, of various winsorized estimators are calculated in Sections 4 and 5. Section 4 derives analytic large sample approximations to the efficiency of Searls' estimator using the theory of extreme order statistics while Section 5 compares, in a Monte Carlo study, estimators for reducing the impact of large data values.

2. SAMPLING PROPERTIES OF THE WINSORIZED MEAN

This section studies winsorized means for data drawn from either a continuous or a discrete distribution. Several families of continuous distributions are available to model positive skewed data. One has the Weibull family, $F_\alpha(x) = 1 - \exp(-(x/\beta)^{1/\alpha})$ for $x > 0$, the log-normal family, $F_\nu(x) = \Phi(\log(x/\beta)/\nu)$ for $x > 0$, and the Pareto family, $F_\gamma(x) = 1 - (1 + x/\beta)^{-\gamma}$ for $x > 0$, where β is a positive scale parameter and α , ν , and γ are positive shape parameters. Discrete skewed distributions arise in survey sampling. Let $\{y_1, \dots, y_N\}$ represent the values of the

¹ Louis-Paul Rivest and Daniel Hurtubise, Département de mathématiques et de statistique, Université Laval, Cité Universitaire, Québec, Canada, G1K 7P4.

variable of interest for the N units of a population to be sampled. If a simple random sample with replacement is drawn, then one can take $F(x) = \sum I(y_i \leq x)/N$ as the underlying distribution where $I(\cdot)$ represents the indicator function. In pps sampling, *i.e.*, sampling with replacement and with probabilities given by $\{p_i, i = 1, \dots, N\}$, one would take $F(x) = \sum p_i I(y_i/(Np_i) \leq x)$. The standard estimator of \bar{y} under pps sampling,

$$\bar{y}_s = \frac{1}{n} \sum_s \frac{y_i}{Np_i}$$

can then be regarded as the mean of a random sample of size n drawn from distribution F . Fuller (1991) provides examples of survey data having skewed distributions.

Let X_1, X_2, \dots, X_n denote a sample drawn from $F(x)$. In pps sampling, one would have $X_i = y_i/(Np_i)$ where p_i and y_i are the selection probability and the value of the y -variable for the i -th unit selected in the sample. The population mean μ is to be estimated by a winsorized mean,

$$\bar{X}_R = \frac{1}{n} \sum_1^n \min(X_i, R) = \bar{X} - \frac{1}{n} \sum_{i=1}^n \max(X_i - R, 0), \quad (2.1)$$

where \bar{X} is the mean of the X_i 's. The expectation of \bar{X}_R is equal to

$$E(\bar{X}_R) = \mu - \int_R^\infty (x - R) dF(x) = \mu - \int_R^\infty \int_R^x dy dF(x).$$

Changing the order of integration in the above integral proves that $E(\bar{X}_R) = \mu + B(\bar{X}_R)$ where

$$B(\bar{X}_R) = - \int_R^\infty [1 - F(x)] dx \quad (2.2)$$

is the bias of the winsorized mean.

By (2.1), an expression for the variance of \bar{X}_R is

$$n \text{Var}(\bar{X}_R) = \sigma^2 - 2 \text{cov}[X_1, \max(X_1 - R, 0)] + \text{Var}[\max(X_1 - R, 0)]$$

where X_1 is the first random variable in the sample and σ^2 is the variance of $F(x)$. Manipulations similar to those yielding (2.2) show that

$$E[\max(X_1 - R, 0)^2] = 2 \int_R^\infty (x - R) [1 - F(x)] dx,$$

and

$$E[\max(X_1 - R, 0)X_1] =$$

$$2 \int_R^\infty (x - R) [1 - F(x)] dx - RB(\bar{X}_R).$$

Thus

$$\text{Var}(\bar{X}_R) =$$

$$\frac{1}{n} \left\{ \sigma^2 - 2 \int_R^\infty (x - \mu) [1 - F(x)] dx - B^2(\bar{X}_R) \right\},$$

and

$$\begin{aligned} \text{MSE}(\bar{X}_R) &= \frac{\sigma^2}{n} - \frac{2}{n} \int_R^\infty (x - \mu) [1 - F(x)] dx \\ &\quad + \frac{n-1}{n} B^2(\bar{X}_R). \end{aligned} \quad (2.3)$$

Searls (1966) showed that the mean square error of \bar{X}_R has a unique minimum which can be obtained by equating the derivative, with respect to R , of $\text{MSE}(\bar{X}_R)$ to 0. This yields the following equation for the optimal winsorization constant $R(F, n)$,

$$\frac{R - \mu}{n - 1} - \int_R^\infty [1 - F(x)] dx = 0. \quad (2.4)$$

This is equivalent to equation (14) in Searls (1966). In the remainder of this work, \bar{X}_R denotes the optimal winsorized mean obtained with the winsorization constant $R(F, n)$ which solves (2.4). Observe that the optimal cut-off point $R(F, n)$ is location and scale equivariant, *i.e.*, if $G(x) = F[(x - b)/a]$, then $R(G, n) = aR(F, n) + b$.

A general algorithm for solving (2.4) is easily constructed. First observe that as a function of R , the left hand side of equation (2.4) is increasing and concave in R since its derivative, $1/(n - 1) + 1 - F(R)$, is positive and decreasing. Therefore, the Newton-Raphson algorithm (Thisted 1988, 164-167) given by

$$R_{j+1} = R_j - \frac{(R_j - \mu) - (n - 1) \int_{R_j}^\infty [1 - F(x)] dx}{1 + (n - 1) [1 - F(R_j)]}, \quad (2.5)$$

with $R_0 = 2\mu$ as starting value converges smoothly to the solution of (2.4). For discrete distributions the computations are easily implemented by noting that

$$\int_R^\infty [1 - F(x)] dx = E[\max(X - R, 0)].$$

Exact calculations of the optimal cut-off points $R(F, n)$ were carried out for the Weibull, the log-normal, and the Pareto families for samples of size s ranging between 5 and 200. Three distributions, corresponding to coefficients of variation (CV) of 1, 2, and 4, were considered in each family except for the Pareto family where only coefficients of variation of 2 and 4 were considered. The CV measures the skewness of a distribution, with large CVs corresponding to heavy skewness. The corresponding parameter values are given in Table 1.

Table 1

Parameter values of the distributions for which optimal cut-off values $R(F, n)$ were evaluated

CV	Weibull(α)	Log-normal(ν)	Pareto(γ)
1	1	0.83	—
2	1.84	1.27	2.67
4	2.87	1.68	2.13

For each distribution and each sample size, the optimal cut-off point was calculated using algorithm (2.5). Figure 1 presents the expected number of winsorized observations, $m(F, n) = n\{1 - F[R(F, n)]\}$ as a function of n while

the corresponding efficiencies are reported in Figure 2. The efficiency of \bar{X}_R is defined as $\text{Var}(\bar{X})/\text{MSE}(\bar{X}_R)$.

In Figure 1 the expected number of winsorized data values under the optimal scheme is, for most skewed distributions, close to 1 even for large sample sizes. Approximating this number by a Poisson distribution with parameter $m(F, n)$ shows there is a non-negligible probability that, under the optimal winsorization scheme, none of the data points is winsorized. This probability increases with the skewness of the distribution since $m(F, n)$ decreases with the CV. Thus, in samples from a highly skewed distribution, it is not always appropriate to winsorize the largest values. Such values should be winsorized only when they are large. As expected, in Figure 2, the largest gains in efficiency are obtained when the skewness is heavy. Therefore monitoring the two or three largest data values in a sample and curtailing their impact when these values are large is the key to a good winsorization strategy.

Figure 1 shows that the expected number of winsorized data values $m(F, n)$ decreases with the skewness of the distribution. This observation can be turned into a rigorous mathematical result. To this end, random variable Y is said to be more skewed than random variable X if Y has the same distribution as $\psi(X)$ where $\psi(x)$ is a convex function

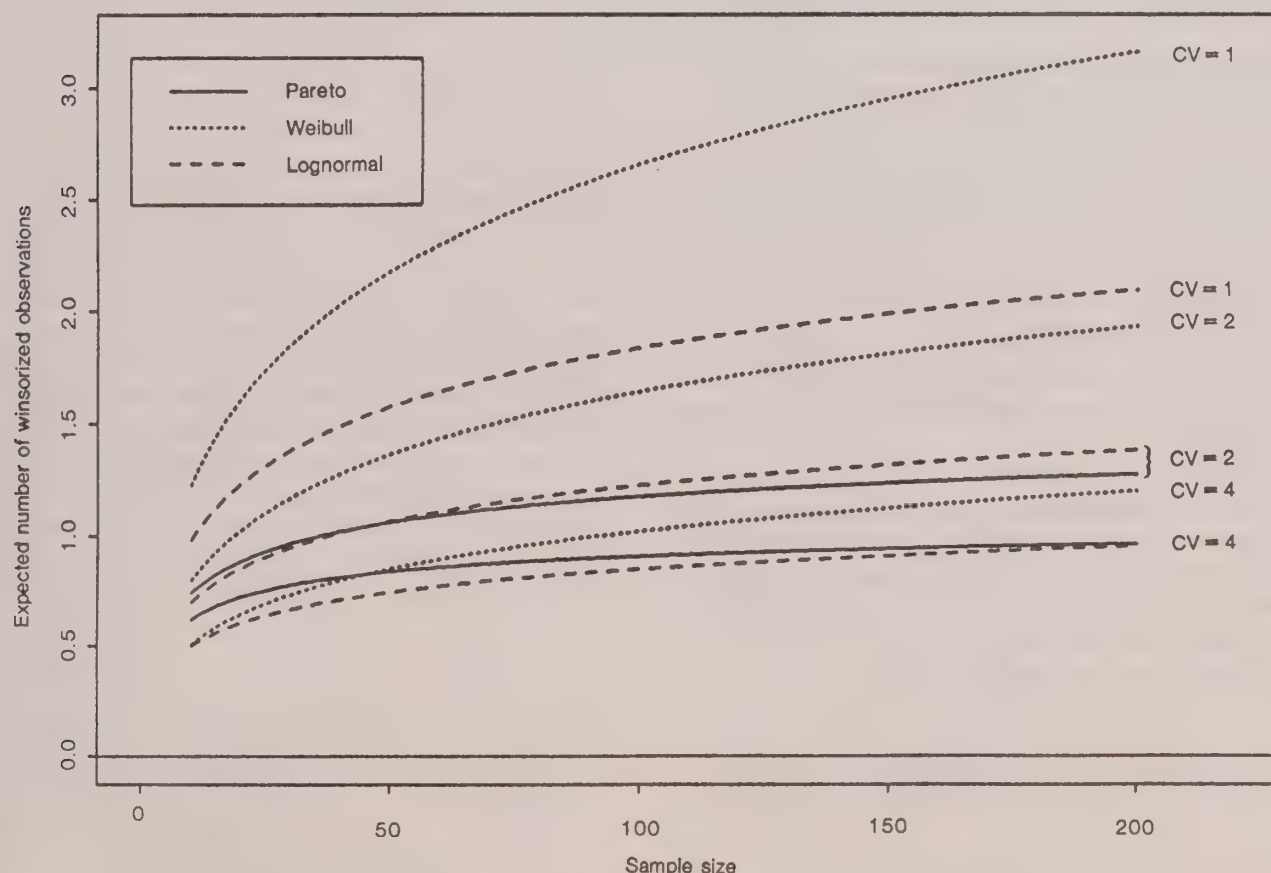


Figure 1. Expected number of winsorized observations for simple and stratified random sampling.

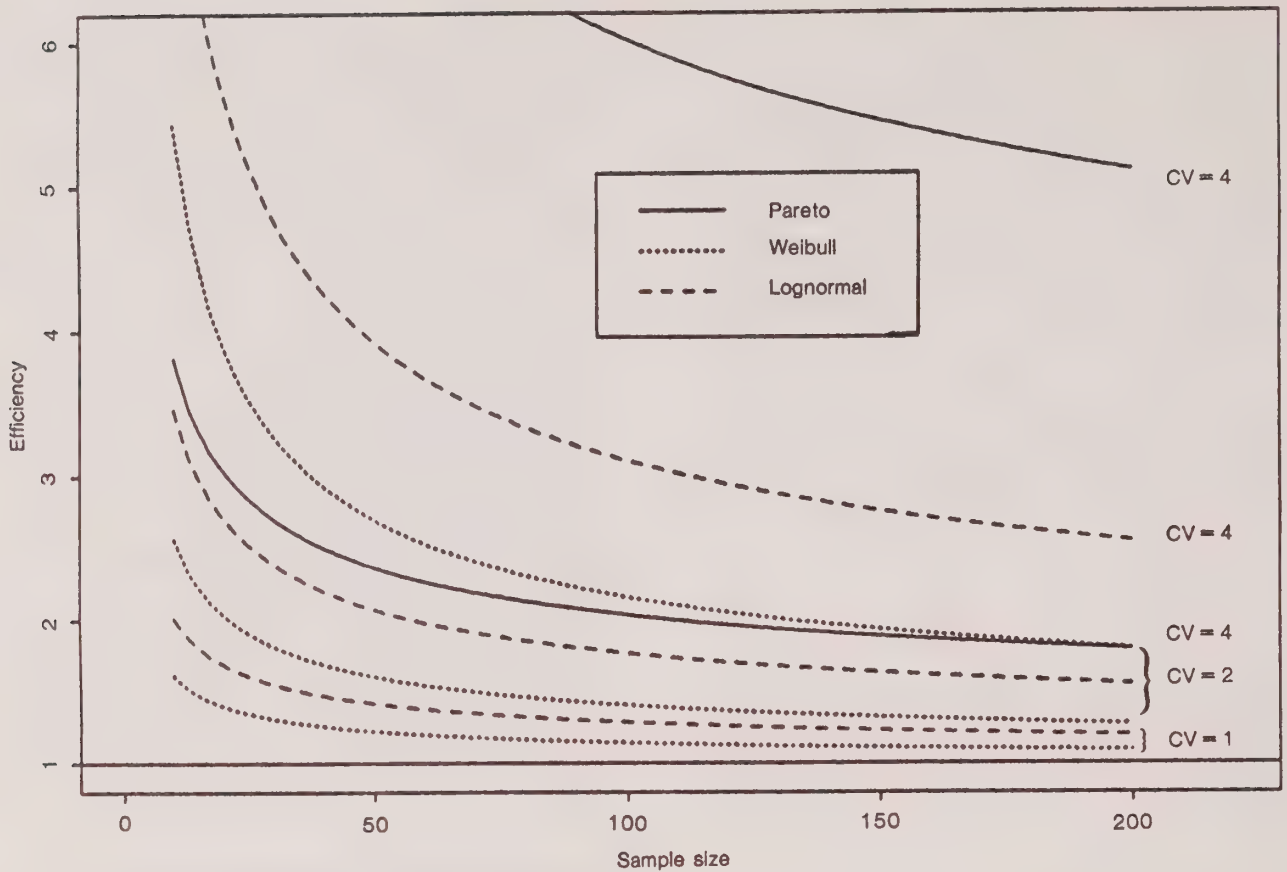


Figure 2. Efficiency of Searls winsorized mean.

of x . Under this definition X^2 is, as should be expected, more skewed than X . This notion of skewness corresponds to the convex partial ordering of van Zwet (Barlow and Proschan 1981). With this definition of skewness, one has the following proposition which is proved in the Appendix together with Propositions 2 and 3.

Proposition 1 If Y is more skewed than X then $m(F_X, n) > m(F_Y, n)$ where F_X and F_Y are the distributions of X and Y respectively.

The results of this section also apply to simple random sampling without replacement. For this design the mean square error of \bar{X}_R is given by formula (2.3) with n replaced by $n/(1-f)$ where f is the sampling fraction. Algorithm (2.5), with n divided by $(1-f)$, can be used for calculating optimal cut-off values for without replacement simple random sampling.

3. WINSORIZATION IN STRATIFIED SAMPLING

There are many ways to generalize Searls' winsorization strategy to stratified sampling. In this section each stratum has its own cut-off value. Let R_h be the cut-off value in

stratum h . The optimal values of R_1, R_2, \dots, R_L , where L is the number of strata, are the ones that minimize the mean square error of $\bar{X}_R = \sum W_h \bar{X}_{Rh}$, where $\bar{X}_{Rh} = \sum \min(X_{hi}, R_h)/n_h$, $W_h = N_h/N$ and N_h is the size of stratum h and $N = \sum N_h$. An algorithm for determining these optimal cut-off values is proposed in this section.

Let $F_h(x)$, for $h = 1, \dots, L$ be the distribution of X in stratum h , and μ_h and σ_h^2 be the mean and the variance of F_h . The derivation of the mean square error of \bar{X}_R , under with replacement stratified random sampling, follows that presented in Section 2, it gives

$$\text{MSE}(\bar{X}_R) = \sum_{h=1}^L \frac{W_h^2}{n_h} \left(\sigma_h^2 - 2 \int_{R_h}^{\infty} (x - \mu_h) [1 - F_h(x)] dx - B^2(\bar{X}_{Rh}) \right) + \left(\sum_{h=1}^L W_h B(\bar{X}_{Rh}) \right)^2 \quad (3.1)$$

where $B(\bar{X}_{Rh})$ is the bias of \bar{X}_{Rh} as an estimator of μ_h

$$B(\bar{X}_{Rh}) = - \int_{R_h}^{\infty} [1 - F_h(x)] dx.$$

Taking the partial derivatives with respect to R_h , $h = 1, \dots, L$ yields the following equations for the optimal values:

$$\frac{W_h}{n_h} [R_h - \mu_h - B(\bar{X}_{Rh})] = - \sum_{h=1}^L W_h B(\bar{X}_{Rh}), \quad (3.2)$$

for $h = 1, \dots, L$.

There is no simple way to solve (3.2). An approximate solution can be obtained by noting that $B(\bar{X}_{Rh})/n_h$ is, for all values of h , usually small as compared to the other terms. Dropping these terms leads to

$$\frac{W_h}{n_h} (R_h - \mu_h) = - \sum_{h=1}^L W_h B(\bar{X}_{Rh}), \quad (3.3)$$

for $h = 1, \dots, L$. The solutions to (3.3) overestimate slightly the optimal values satisfying (3.2) since at these solutions the partial derivatives of (3.1) are all positive and since these partial derivatives are increasing functions of R_h , for $h = 1, \dots, L$. Thus by solving (3.3) to estimate the cut-off values one does not run the risk of winsorizing too many data values. Equations (3.3) imply that $R_h = \mu_h + n_h R / (n W_h)$ where R is some positive constant. A simple equation for R is obtained by changing variable $y = n W_h (x - \mu_h) / n_h$ in the integrals for $B(\bar{X}_{Rh})$, $h = 1, \dots, L$ where $n = \sum n_h$. This gives

$$- \sum_{h=1}^L W_h B(\bar{X}_{Rh}) = \frac{R}{n} = \int_R^{\infty} [1 - F(y)] dy = -B(\bar{X}_R), \quad (3.4)$$

where $F(y) = \sum n_h F_h[\mu_h + n_h y / (n W_h)] / n$. Equation (3.4) is easily solved using algorithm (2.5) proposed in Section 2 for the single sample case. Therefore simple approximations for Searls' optimal cut-off values in stratified sampling are easily calculated.

Since the distribution F defined above has a zero expectation, the mean square error of the stratified winsorized mean obtained by solving (3.3) is equal to:

$$\begin{aligned} \text{MSE}(\bar{X}_R) &= \frac{1}{n} \\ &\left(\sigma_F^2 - 2 \int_R^{\infty} y [1 - F(y)] dy - B(\bar{X}_R)^2 \right) + B(\bar{X}_R)^2 \\ &+ \left(\frac{1}{n} B(\bar{X}_R)^2 - \sum_{h=1}^L \frac{W_h^2 B^2(\bar{X}_{Rh})}{n_h} \right) \end{aligned} \quad (3.5)$$

where σ_F^2 is the variance of F . The last term of (3.5) is easily shown to be negative or null; it is null when $B(\bar{X}_R) = n W_h B(\bar{X}_{Rh}) / n_h$ for $h = 1, \dots, L$. The variance of the stratified mean, $\bar{X} = \sum W_h \bar{X}_h$, is equal to σ_F^2 / n . Thus a conservative approximation to the efficiency of \bar{X}_R with respect to \bar{X} in stratified sampling is equal to the corresponding efficiency for a random sample of size n drawn from F . Note also that $n[1 - F(R)]$ represents the expectation of the total number of winsorized data points in the L strata.

The optimal winsorization scheme obtained by solving (3.3) has a simple form for many allocation rules. Under proportional allocation, *i.e.*, $n_h = n W_h$ for $h = 1, \dots, L$, one gets $R_h = \mu_h + R$. Under Neyman optimal allocation, with $n_h = n W_h \sigma_h / (\sum W_h \sigma_h)$ where σ_h is stratum h 's standard deviation, one gets $R_h = \mu_h + \sigma_h R / (\sum W_h \sigma_h)$. If in addition, the distributions of X within the strata are equal up to a change in location and scale, *i.e.*, $F_h = F_0[(x - \mu_h) / \sigma_h]$ for some distribution F_0 , then $F(x) = F_0[x / (\sum W_h \sigma_h)]$. In this case the characteristics of optimal winsorized means in stratified sampling and in simple random sampling are the same. Thus Figure 1 presents the expected total number of winsorized data points in the L strata as a function of the total sample size n , under Neyman allocation, when F_0 is one of the distributions of Table 1. Figure 2 gives the corresponding efficiencies.

The results of this section are easily generalized to without replacement stratified sampling by replacing n_h by $n_h / (1 - f_h)$ throughout the calculations. The derivation of optimal cut-off values for stratified pps sampling is easily carried out by taking $F_h(x) = \sum p_{hi} I(y_{hi} / (N_h p_{hi}) \leq x)$ where p_{hi} denotes the selection probability for unit the i -th unit of stratum h .

4. LARGE SAMPLE APPROXIMATIONS TO THE EFFICIENCY OF THE WINSORIZED MEAN

For most distributions, equation (2.3) defining the optimal cut-off does not have an explicit solution. This section derives closed form approximations to this solution using the theory of extreme order statistics. This will permit the derivation of explicit approximations to the efficiency of the optimal winsorized mean. Searls' optimal winsorization strategy will then be compared to a simple non parametric winsorization scheme where the largest order statistic is replaced by the second largest (Rivest 1994).

The form of the approximation to $R(F, n)$ depends on the limiting distribution, as the sample size n goes to infinity, of the largest order statistic suitably normalized. For distributions whose support is the positive axis, there are only two possible limiting distributions which are given by Galambos (1987, p. 53-54)

$$H_{1,\alpha}(x) = \exp(-x^{-\alpha}) \quad \text{for } x > 0 \quad \text{and } \alpha > 0$$

and

$$H_{3,0}(x) = \exp[-\exp(-x)] \quad \text{for } x \text{ in } R.$$

For many distributions used for the statistical analysis of positive random variables, for example the Weibull and the log-normal families, the sample maximum suitably normalized converges to $H_{3,0}(x)$. Distributions whose sample maxima converge to $H_{1,\alpha}(x)$ for some $\alpha > 0$ have heavy tails. For such distributions $1 - F(x)$ goes to 0 at a rate of $O(x^{-\alpha})$. The Pareto and the F distributions are in this class.

Distributions whose sample maxima converge to $H_{3,0}(x)$ are considered first. The following characterization is due to von Mises (1964): the sample maximum of a twice differentiable distribution $F(x)$ converges to $H_{3,0}(x)$ if, as x goes to ∞ ,

$$\lim_x \frac{g'(x)}{g^2(x)} = 0 \quad (4.1)$$

where $f(x)$ is the density of F , $g(x) = f(x)/[1 - F(x)]$ is the failure rate of F , and g' is the derivative of g . An approximation to winsorization constant $R(F, n)$ for this class of distributions is presented next.

Proposition 2 If $F(x)$ is such that (4.1) holds and if, for large values of x , it satisfies:

- i) $xg(x)$ increases;
 - ii) $xg'(x)/g(x)$ is less than some positive constant c ;
- then the optimal winsorization constant $R(F, n)$ satisfies

$$R(F, n) =$$

$$F^{-1}\left(1 - \frac{g[F^{-1}(1 - 1/n)]F^{-1}(1 - 1/n)[1 + o(1)]}{n}\right);$$

and $m(F, n) = g(F^{-1}(1 - 1/n))F^{-1}(1 - 1/n)[1 + o(1)]$. Furthermore, the mean squared error of Searls' winsorized mean is approximately equal to

$$\text{MSE}(\bar{X}_R) \approx \frac{\sigma^2}{n} - \frac{R(F, n)^2}{n^2}.$$

In the Weibull family, $F_\alpha^{-1}(1 - t) = [-\log(t)]^\alpha$, $g(x) = x^{1/\alpha-1}/\alpha$. The hypotheses of Proposition 2 are met and $m(F_\alpha, n)$, the expected number of winsorized observations in a large Weibull sample, is $\log(n)[1 + o(1)]/\alpha$ which goes to ∞ as n increases. Figure 1 suggests that the convergence is very slow, especially for large coefficients of variation.

Now consider distributions whose sample maxima converge to $H_{1,\alpha}(x)$. This class of distributions has been characterized by Gnedenko (1962): the sample maximum of F converges to $H_{1,\alpha}(x)$ if one can write

$$1 - F(x) = L(x)/x^\alpha \quad (4.2)$$

where as x goes to ∞ , $L(x)/L(kx)$ converges to 1, for any constant k . Note that for F to have a finite second moment, one needs $\alpha > 2$ in (4.2). The Pareto distribution satisfies (4.2) with $\alpha = \gamma$.

Proposition 3 If F satisfies (4.2) with parameter α where $\alpha > 2$, then as n goes to infinity, $R(F, n) = F^{-1}[1 - (\alpha - 1)/n][1 + o(1)]$, i.e., $m(F, n) \approx \alpha - 1$. Furthermore,

$$\text{MSE}(\bar{X}_R) \approx \frac{\sigma^2}{n} - \frac{\alpha R(F, n)^2}{n^2(\alpha - 2)}.$$

For distributions satisfying (4.2) a finite number of data points are on average winsorized as the sample size goes to ∞ . To some extent, this can be seen in Figure 1 where the curves of $m(F_\gamma, n)$ for the Pareto distribution have $m(F_{2.33}, n) = 1.33$, and $m(F_{2.67}, n) = 1.67$ as asymptotes.

Propositions 2 and 3 shed some light on the estimation of the optimal cut-off value. When F is unknown, a possible estimator for $R(F, n)$ is the value that minimizes an estimator of the mean square error of \bar{X}_R . This leads to

$$\frac{R - \bar{X}}{n - 1} = \frac{1}{n} \sum_{i=1}^n \max(X_i - R, 0) \quad (4.3)$$

as an estimating function for R . This procedure is questionable when the underlying distribution is highly skewed, i.e., when F satisfies the assumption of Proposition 3. On average, there will only be $\alpha - 1$ non-null terms in the right hand side of equation (3). Thus \hat{R} will, on average be determined by the $\alpha - 1$ largest data values and the sample maximum will have the largest influence on \hat{R} . This will make \hat{R} highly unstable and, considering the findings of Figure 1, the second largest sample order statistic should be a better estimator of $R(F, n)$ than the solution of (3.3). This is exemplified in the Monte Carlo simulations of Section 5.

Table 2 compares approximations to the bias and to the mean square error of Searls' winsorized mean \bar{X}_R to those of the once winsorized mean \bar{X}_1 obtained by taking the cut-off value R equal to the second largest observation. Rivest (1994) shows this choice of cut-off value yields the optimal non-parametric winsorized mean. He also derives the large sample approximations for the bias and the mean square error of \bar{X}_1 appearing in Table 2. The corresponding expressions for \bar{X}_R are taken in Propositions 2 and 3.

Table 2

Approximations to the bias and to the mean square error of the once winsorized mean \bar{X}_1 and of Searls' optimal winsorized mean, \bar{X}_R , for the Weibull and for the Pareto distribution ($\Gamma(\cdot)$ stands for the gamma function)

		WEIBULL		PARETO	
\bar{X}_R	MSE	$\frac{\sigma^2}{n} - \frac{(\log n)^{2\alpha}}{n^2}$		$\frac{\sigma^2}{n} - \frac{\gamma}{(\gamma - 2)(\gamma - 1)^{2/\gamma} n^{2-2/\gamma}}$	
	bias	$-\frac{(\log n)^\alpha}{n}$		$-\frac{1}{(\gamma - 1)^{1/\gamma} n^{1-1/\gamma}}$	
\bar{X}_1	MSE	$\frac{\sigma^2}{n} - \frac{2\alpha(\alpha - 1)(\log n)^{2\alpha-2}}{n^2}$		$\frac{\sigma^2}{n} - \frac{2\Gamma(1 - 2/\gamma)}{\gamma(\gamma - 1)n^{2-2/\gamma}}$	
	bias	$-\frac{\alpha(\log n)^{\alpha-1}}{n}$		$-\frac{\Gamma(1 - 1/\gamma)}{\gamma n^{1-1/\gamma}}$	

In Table 2 the mean square error of \bar{X}_R is much smaller than that of \bar{X}_1 . Indeed, for the Weibull distribution the large sample efficiency of \bar{X}_R with respect to \bar{X}_1 is equal to that of \bar{X}_R with respect to \bar{X} . Thus non-parametric winsorization reduces the mean square error of estimators of the mean of a skewed population however further reductions in mean square error can be obtained if information concerning the underlying distribution is available. This is illustrated in the Monte Carlo comparisons presented in the next section.

The results of this section apply to stratified sampling. For this design, the large sample solution to equation (3.4) is determined by the stratum with the most skewed distribution. If F_1 is the most skewed distribution then $nW_1R(F_1, n_1)/n_1$ is an approximate solution to (3.4) where an approximation to $R(F_1, n_1)$ is found in Proposition 2 or in Proposition 3 depending on the tail of F_1 . In this case only data points in stratum 1 are winsorized in large stratified samples. Searls' winsorized mean is then equal to W_1 times the optimal winsorized mean for stratum one plus a weighted sum of the sample means in the other strata.

5. MONTE CARLO COMPARISONS OF ESTIMATORS OF THE MEAN OF A SKEWED DISTRIBUTION

This section presents Monte Carlo comparisons of the mean square error and of the biases of five estimators of the mean of population CHICKEN of Fuller (1991). This population has 2000 units; its coefficient of variation is

4.46. Further numerical comparisons of the five estimators considered in this section for other distributions, either finite or infinite, are presented in Rivest (1993a and b).

The five estimators under consideration are:

- Searls' winsorized estimator, \bar{X}_R , calculated as if the underlying distribution was known;
- A winsorized estimator where the cut-off value is set equal to the second largest data value of an auxiliary sample of size $2n$; this is an instance where limited auxiliary information concerning the underlying distribution F is available (in the Monte Carlo simulations each simulated sample had its own auxiliary sample);
- The once winsorized mean, \bar{X}_1 , introduced in Section 4;
- A winsorized estimator where R is estimated from the sample by solving equation (4.3);
- Fuller preliminary test estimator with $j = 3$ (i.e., the numerator of the preliminary test involves the three largest observations), T (the total number of data points involved in the preliminary test) equal to $[4n^{1/2} - 10]$ and K_3 , the cut-off value equal to 3.5. A detailed description of this estimator appears in Fuller (1991) and in Rivest (1993a and b). This estimator curtails the largest data values only when a test statistic for detecting extreme data values is significant.

The biases and the efficiencies of \bar{X}_R were calculated exactly. For the other estimators, the biases and the efficiencies presented in Figures 1 and 2 were obtained in Monte Carlo simulations based on 100,000 repetitions.

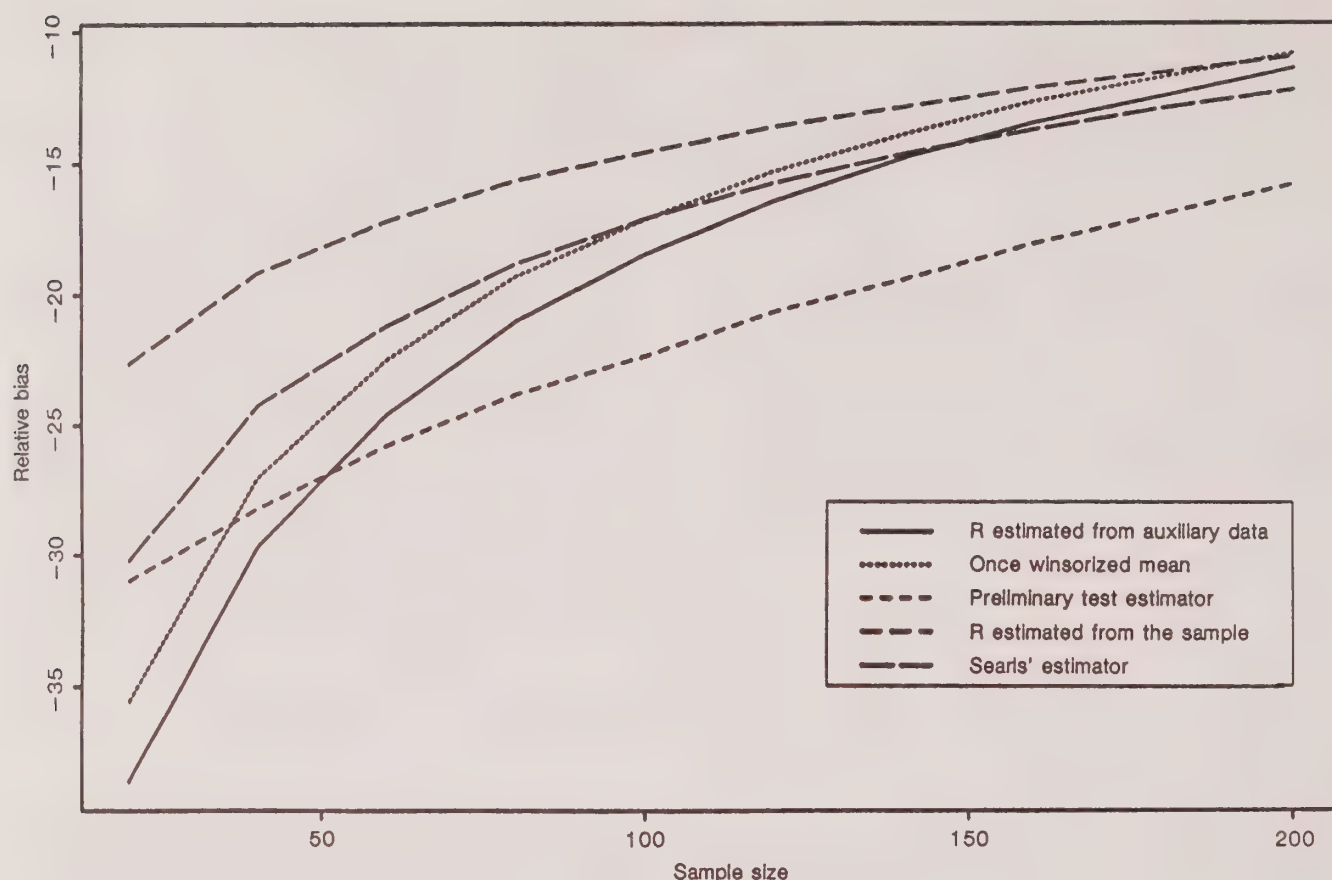


Figure 3. Relative bias of five estimators for the mean of CHICKEN.

Figure 3 indicates that the biases of winsorized estimators are important, even in large samples. Several interesting conclusions can be drawn from Figure 4. First, as expected from Table 2 Searls' estimator is much more efficient than the once winsorized mean. Estimating the optimal cut-off value using limited auxiliary information is highly efficient. This holds true as long as the study variable can be modeled by a superpopulation distribution having a finite variance, see Rivest (1993a) for further discussions. In a sampling context, the auxiliary samples could be data from previous surveys standardized to account for possible changes over time in the distribution of the variable under study.

Among the three estimators of Figure 4 that do not rely on auxiliary information, Fuller estimator is the best. This is in agreement with the simulation results of Fuller (1991). Estimating the cut-off value by minimizing an estimate of the mean square error does poorly especially in small samples. Thus, as shown in Section 4, the resulting estimator is highly sensitive to the wild data values that sometimes appear in small samples. This estimator is not recommended.

6. CONCLUSIONS

Many strategies can be used to accommodate the large values that sometimes arise in surveys. If auxiliary information, such as census data, is available then one can use Searls' estimator in either simple random sampling, stratified sampling, or pps sampling. Since the cut-off values are fixed constant mean square error estimators can be derived from formulae (2.3) and (3.1).

When extra information is not available, the once winsorized mean and Fuller preliminary test estimator can be used. Research is now under way to generalize these estimators to stratified designs. An estimator for the mean square error of the once winsorized mean is proposed in Rivest (1994),

$$v(\bar{X}_1) = \frac{1}{n} S^2 - \frac{1}{n^2} (X_n + X_{n-1} - 2\bar{X}_1) \\ (X_n - 3X_{n-1} + 2X_{n-2})$$

where S^2 denotes the variance of the X -sample and $X_n > X_{n-1} > X_{n-2}$ denote the three largest data values in

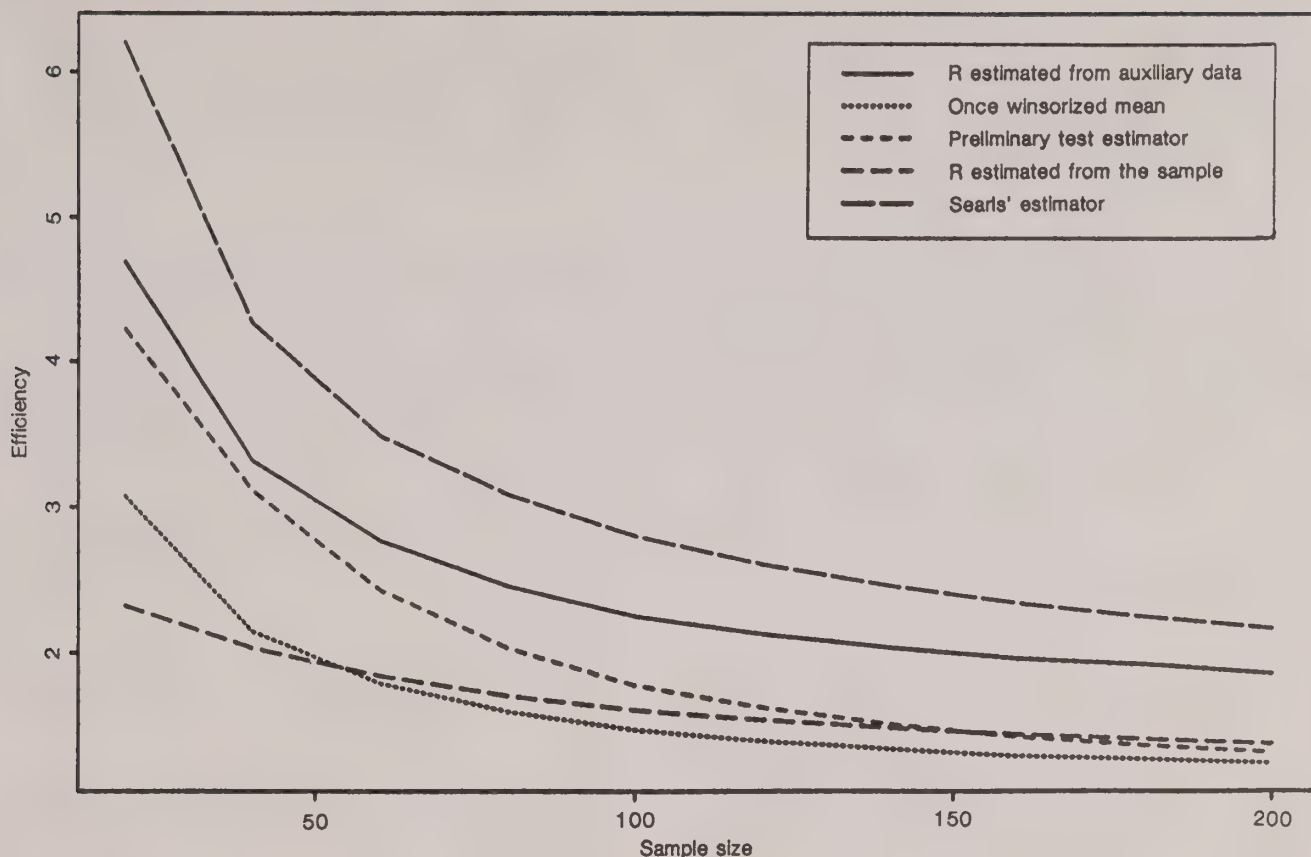


Figure 4. Efficiency of five estimators for the mean of CHICKEN.

that sample. This estimator has a small bias in infinite populations. However the coverage of the standard confidence interval $\bar{X}_1 \pm z_{1-\alpha/2} \sqrt{v(\bar{X}_1)}$ is often well below the nominal $100(1 - \alpha)\%$ level especially when the underlying distribution is skewed. Further research is needed to obtain reliable confidence intervals for estimators of the mean of skewed populations.

ACKNOWLEDGEMENTS

This research was supported by the Natural Science and Engineering Research Council and by the Fond pour la formation des chercheurs et l'aide à la recherche of Québec.

APPENDIX 1

Proof of Proposition 1 The assumption that Y is more skewed than X implies that there exists a convex function ψ such that $\psi(X)$ and Y have the same distribution. Let R denote $R(F_X, n)$. To prove the result, it suffices to show that $\psi(R) < R(F_Y, n)$. This is equivalent to

$$\frac{\psi(R) - E(Y)}{n-1} < \int_{\psi(R)}^{\infty} [1 - F_Y(x)] dx. \quad (\text{A.1})$$

By Jensen's inequality, $E(Y) = E[\psi(X)] > \psi[E(X)]$. Thus using (2.3), the left hand side of (A.1) is less than or equal to

$$\frac{R - E(X)}{n-1} \frac{\psi(R) - \psi[E(X)]}{R - E(X)} < \frac{R - E(X)}{n-1} \psi'(R) = \int_R^{\infty} [1 - F_X(y)] dy \cdot \psi'(R)$$

where ψ' is the derivative of ψ . Since ψ' is increasing, the left hand side of the above inequality is less than or equal to:

$$\int_R^{\infty} \psi'(y) [1 - F_X(y)] dy = \int_{\psi(R)}^{\infty} [1 - F_Y(x)] dx.$$

This shows that (A.1) holds.

Proof of Proposition 2 The following result obtained by applying Theorems 2.7.5 and 2.7.11 of Galambos (1987) to the distribution $F(z^{p+1})$ is used extensively. If the sample maxima of distribution $F(x)$ converges to $H_{3,0}(x)$, then all the moments of F exist and

$$\int_x^\infty y^p [1 - F(y)] dy \sim \frac{[1 - F(x)] x^p}{g(x)} \quad (\text{A.2})$$

where $g(x) \sim h(x)$ means that $g(x)/h(x)$ converges to 1 as x goes to infinity. Using (A.2), $R(F, n)$ is obtained by solving

$$\frac{R - \mu}{n - 1} = \frac{1 - F(R)}{g(R)} (1 + o(1)).$$

Let $R = F^{-1}(1 - a/n)$, then, up to $(1 + o(1))$, the above equation becomes

$$a = g \left[F^{-1} \left(1 - \frac{a}{n} \right) \right] F^{-1} \left(1 - \frac{a}{n} \right). \quad (\text{A.3})$$

Let $a_0 = g[F^{-1}(1 - 1/n)]F^{-1}(1 - 1/n)$ and $a_1 = g[F^{-1}(1 - a_0/n)]F^{-1}(1 - a_0/n)$. Since for large values of x , $g(x)$ is increasing, $a_0 > a_1$ and the solution to (A.3) belongs to the interval (a_1, a_0) . In order to prove the result, one has to show that a_1/a_0 converges to 1 as n goes to ∞ .

Since $g(x) = f(x)/[1 - F(x)]$, one can write

$$a_0 = \exp \left[\int_{F^{-1}(1-a_0/n)}^{F^{-1}(1-1/n)} g(t) dt \right] = \exp \left[a_0 - a_1 - \int_{F^{-1}(1-a_0/n)}^{F^{-1}(1-1/n)} tg'(t) dt \right],$$

where the second expression is obtained by integrating by parts. Since $tg'(t)/g(t)$ is less than c , one has $a_0 > \exp(a_0 - a_1)a_0^{-c}$. If a_1/a_0 does not converge to 1, say $a_1/a_0 < 1 - \epsilon < 1$ for an infinite sequence of sample sizes, the previous inequality implies that $a_0^{1+c} > \exp(a_0\epsilon)$. This is a contradiction since a_0 tends to ∞ as n becomes large. The approximation for $\text{MSE}(\bar{X}_R)$ is obtained by using (A.2) with $p = 2$.

Proof of Proposition 3 If the sample maxima of distribution $F(x)$ converges to $H_{1,\alpha}(x)$ then F satisfies the following properties (Feller 1971, p. 281):

$$\int_x^\infty y^p [1 - F(y)] dy \sim \frac{[1 - F(x)] x^{p+1}}{\alpha - p - 1} \quad (\text{A.4})$$

for any p such that $\alpha - p - 1 \geq 0$. By (A.4), $R(F, n)$ is obtained by solving $F(R) = 1 - [\alpha - 1 + o(1)]/n$. This leads to the approximation for $R(F, n)$. To derive the approximation for $\text{MSE}(\bar{X}_R)$, one applies (A.4) with $p = 1$.

REFERENCES

- BARLOW, R.E., and PROSCHAN, F. (1981). *Statistical Theory of Reliability and Life Testing*. Silver Spring MD: To Begin With.
- CHAMBERS, R.L., and KOKIC, P.N. (1993). Outlier robust sample survey inference. *Bulletin of the International Statistical Institute. Proceedings of the 49th session*, book 2, 54-72.
- ERNST, L.R. (1980). Comparison of estimators of the mean which adjust for large observations. *Sankhyā C*, 42, 1-16.
- FELLER, W. (1971). *An Introduction to Probability Theory and its Applications*. Volume II. Second Edition. New York: Wiley.
- FULLER, W.A. (1991). Simple estimators for the mean of skewed populations. *Statistica Sinica*, 1, 137-158.
- FULLER, W.A. (1993). Estimators for long-tailed distributions. *Bulletin of the International Statistical Institute. Proceedings of the 49th session*, book 2, 39-54.
- GALAMBOS, J. (1987). *The Asymptotic Theory of Extreme Order Statistics*. Second edition. Malabar FL: Krieger.
- GNEDENKO, B.V. (1962). *The Theory of Probability*. New York: Chelsea.
- GLASSER, G.J. (1962). On the complete coverage of large units in a statistical study. *Review of the International Statistical Institute*, 30, 28-32.
- HICKS, S., and FETTER, M. (1993). An evaluation of robust estimation techniques for improving estimates of total hogs. *Proceedings of the International Conference on Establishment Surveys*. Alexandria, Virginia: American Statistical Association, 385-389.
- HIDIROGLOU, M.A. (1987). The construction of a self-representing stratum of large units in survey design. *The American Statistician*, 40, 27-31.
- HUBER, P.J. (1981). *Robust Statistics*. New York: John Wiley.
- LEE, H. (1994). Outliers in Survey Data. Statistics Canada paper.
- RIVEST, L.-P. (1994). Some sampling properties of winsorized means for skewed distributions. *Biometrika*, 81, 373-384.
- RIVEST, L.-P. (1993a). Winsorization of survey data. *Bulletin of the International Statistical Institute. Proceedings of the 49th session*, book 2, 73-89.
- RIVEST, L.-P. (1993b). Winsorization of survey data. *Proceedings of the International Conference on Establishment Surveys*. Alexandria, Virginia: American Statistical Association, 396-401.
- SEARLS, D.T. (1966). An estimator which reduces large true observations. *Journal of the American Statistical Association*, 61, 1200-1204.
- THISTED, R.A. (1988). *Elements of Statistical Computing*. New York: Chapman and Hall.
- VON MISES, R. (1964). *Selected Papers of Richard von Mises*. Volume II. Providence: American Mathematical Society.

Design Effects for Correlated ($P_i - P_j$)

LESLIE KISH, MARTIN R. FRANKEL, VIJAY VERMA and NIKO KAČIROTI¹

ABSTRACT

We present empirical evidence from 14 surveys in six countries concerning the existence and magnitude of design effects (defts) for five designs of two major types. The first type concerns $\text{deft}(p_i - p_j)$, the difference of two proportions from a polytomous variable of three or more categories. The second type uses Chi-square tests for differences from two samples. We find that for all variables in all designs $\text{deft}(p_i - p_j) \cong [\text{deft}(p_i) + \text{deft}(p_j)]/2$ are good approximations. These are *empirical* results, and exceptions disprove the existence of mere analytical inequalities. These results hold despite great variations of defts between variables and also between categories of the same variables. They also show the need for sample survey treatment of survey data even for analytical statistics. Furthermore they permit useful approximations of $\text{deft}(p_i - p_j)$ from more accessible $\text{deft}(p_i)$ values.

KEY WORDS: Design effects; Survey sampling; Sampling errors.

1. DESIGN EFFECTS FOR ANALYTICAL STATISTICS

We explore the existence and the magnitudes of design effects for some special analytical statistics based on data from survey samples. The investigation is both methodological and empirical, with data from several different surveys with different variables and from contrasting populations, hence subject to the risks of inconsistent empirical results. We often hear and read that probability sampling, while necessary for descriptive surveys, is not necessary for analytical surveys. In "Four Obstacles to Representation in Analytic Studies" one of us wrote that "In addition to those four real obstacles, we also encounter another, which is more artificial, in the denials of the need for representation" (Kish 1987, Section 2.7). Sampling investigations show that complex probability selections, especially clustered sampling, have no appreciable influence on descriptive statistics (like means and regression coefficients), but can have drastic effects on inferential statistics, like confidence intervals, tests of significance (Kish and Frankel 1974).

Design effects are defined as $\text{deft}^2 = \text{actual variance} / \text{simple random variance of same } n$, both estimated. And values of $\text{deft} > 1$ have been shown for sampling errors not only of means, but also for analytical statistics like differences of means (and Chi square tests), regression coefficients *etc.* It is true that considerable reductions and differences of deft values have been found for some analytical statistics. The differing deft values are not mere necessary mathematical consequences of the sample design, which may be deduced once for all. They have

empirical content and therefore they need to be replicated with empirical investigations (Kish and Frankel 1974; Kish 1987, 7.1; Kish 1965, 14.1-14.2; Rao and Wu 1985; Scott and Holt 1982; Skinner, Holt, and Smith 1989). In this paper we investigate the possible effects and the magnitudes of design effects for a set of related statistics that have not been investigated before. On the contrary, in several statistical papers the absence of design effects was merely assumed by the authors (all justly famous), and apparently passed on by the journal referees, without warning the readers. We shall see if deft is reduced or eliminated for this set of analytical statistics (Cochran 1950; Mosteller 1952; Scott and Seber 1983; Seber and Wild 1993).

Furthermore, we also propose explicitly, as has been implied before, that the existence of considerable values of deft is strong evidence for the need for probability selections. It would be difficult to assume a model of a population distribution where the selection design was unimportant (or uninformative) but produced considerable design effects. The reverse does not hold: absence of design effects is necessary but not sufficient evidence for license to neglect probability selection. This proposition gives added importance to our study, which relates $\text{deft}(p_i - p_j)$ for analytical statistics to $\text{deft}(p_i)$ and $\text{deft}(p_j)$ for two of several categories of the same variable.

Section 2 describes the five related problems (designs) for which sampling errors are described in Section 3. Section 4 discusses the empirical evidence in the tables. Section 5 places our findings in the context of earlier work on defts for subclasses and their differences.

¹ Leslie Kish, ISR, University of Michigan, Ann Arbor MI 48106, U.S.A.; Martin R. Frankel, NORC and City University of New York; Vijay Verma, University of Essex, Colchester, C04 3SQ, U.K.; Niko Kačiroti, Institute of Statistics, Tirana, Albania.

2. SIMILAR STATISTICS FOR FIVE DESIGNS

It has been shown that five designs (problems), of two distinct types, can be treated with the same simple statistics (Kish 1965, Section 12.10). For our empirical and simple presentation we use symbols for sample values (like d , p_i and n_i), even when occasionally capitals for population values would be more appropriate.

The difference of proportions $p_2 - p_0 = n_2/n - n_0/n$ expresses the desired estimate, where $n = n_0 + n_1 + n_2 + \dots + n_k$ is the sample size, with n units selected and weighted equally. Furthermore, under simple random sampling assumptions, the variance of $(p_2 - p_0)$ is $(1 - f)[p_2 + p_0 - (p_2 - p_0)^2]/(n - 1)$.

Type A Comparisons

1. The difference between two categories ($n_2 - n_0$)/ $n = (p_2 - p_0)$ of a polytomy can represent preference between two parties among several (k) in voting surveys, or between two brands of automobiles in market research, or two of several attitudes, opinions, behaviors on one variable, *etc.* The other ($k - 2$) choices are summed into p_1 and disregarded in the difference. (Also treated by Scott and Seber 1983.)
2. Rank values of $-1, 0, +1$ (or $0, 1, 2$ or $c, c + 1, c + 2$) can be assigned to an ordered trichotomous variable without a metric, and viewed as a simple form of the difference of two categories. This form is particularly useful for computations of sampling errors, because all the five designs can use $-1, 0, +1$ for instance as a transformed computing variable.
3. The difference of proportions from two different variables (x and y) may be treated as in (1) and (2). Define as positive in x (or success) only those elements that are positive in x but not in y , so that $n_{10} = n(x_1, y_0)$. Similarly define as positive y the $n_{01} = n(x_0, y_1)$. Then $(n_{10} - n_{01})/n = (p_x - p_y)$ is the net difference in the proportion of positives in x and y . Those that are positives or negatives in both x and y do not count in the differences. Thus we have a case of three categories as in (1) and (2). An example is the difference between the proportions who would "stop all nuclear testing," and those who "want complete nuclear disarmament"; or who would "force Iraq to leave Kuwait" and who would "remove Saddam from power," (Wild and Seber 1993). However, the two categories may also come from two *different surveys* of the same n cases, as in a quality check, or from dual frame observations, or from two waves of a sample. These situations resemble those of (4) and (5).

Type B Comparisons

4. Test-retest and before-after are terms for designs in which the same subjects undergo two observations. Then dichotomous answers $n_2 = n_{10}$ denote the number of negative changes; $n_0 = n_{01}$ the number of positive changes; and $n_{11} + n_{00}$ the sum of the unchanged positives and negatives. Positive and negative answers are respectively denoted here as 1 and 0, and the first and second wave by the order of the subscript. The difference $(n_{10} + n_{11}) - (n_{01} + n_{11}) = n_{10} - n_{01} = n_2 - n_0$ measures the change between positives for the two observations; and $p_2 - p_0 = n_2/n - n_0/n$ measures the change in proportions. (McNemar 1949; Cochran 1950; Mosteller 1952).
5. Matched pairs of n pairs of subjects can also be treated as a generalization of the test-retest design (Mosteller 1952). For example n pairs of randomized subjects may represent experimental versus control treatments; or n pairs of boys versus girls matched on control variables. The statistical treatment $(p_{10} - p_{01})$ of the n pairs of matched subjects is the same as for the n pairs of treatments on the same n subjects (4).

The similarity of statistical treatment for these five designs of two distinct types is convenient, and we present empirical results for both types. "It also has heuristic value that has been overlooked in recent publications (Scott and Seber 1983 and Wild and Seber 1993). The Chi-square test for types 4 and 5 was published early (McNemar 1949; Cochran 1950; Mosteller 1952), and the similarity to the categorical cases 1, 2, 3 was shown" (Kish 1965, 12.10). (Kish was wrong in denoting "trichotomies and matched dichotomies," as "Trinomials and Matched Binomials," which terms refer to IID samples only.)

All of these deal with differences of proportions p_i based on count variables n_i . Extensions to correlated differences $(y_i - y_j)$ for other variables are possible, but not within the scope of our study. Practical examples would include the difference in dollar shares (not only numbers n_i) between two automobile makes from a total of $\sum y_i$ sales.

3. SAMPLING ERRORS AND DESIGN EFFECTS

For simple random samples of size n it can be easily shown (Kish 1965, 12.10) that

$$\text{var}(p_2 - p_0) =$$

$$\left[\frac{(1 - f)n}{(n - 1)} \right] [p_2 + p_0 - (p_2 - p_0)^2]/n.$$

Most of the examples found and shown come from large survey samples, where the $(1 - f)$ can be disregarded. It is worth noting that for the element variance

$$p_2 + p_0 - (p_2 - p_0)^2 = p_2 q_2 + p_0 q_0 + 2p_2 p_0,$$

where the last term $\text{cov}(p_2, p_0) = -p_2 p_0$ represents the covariance arising because p_2 and p_0 are competitive parts of the same sample, rather than proportions from independent samples. The difference of proportions squared $(p_2 - p_0)^2$ will usually be a small correction term, and without it we have the equivalent of the variance $(p_2 + p_0)/n$ of two independent Poisson samples. Furthermore, note that (Kish 1965, 12.10):

The Chi-square test has been applied to some of these problems, treated separately (Cochran 1950; Mosteller 1952; McNemar 1962, p. 225). This is essentially $(n_2 - n_0)^2 / (n_2 + n_0)$ the square of the difference divided by its variance, under the null hypothesis $n_2 = n_0$. It applies the exact theories available for tests of null hypotheses in small samples, including the "Yates correction," all based on the assumption of simple random sampling. However, there are great advantages in treating these problems in large samples as estimated means with proper standard errors. First, instead of being confined to testing null hypotheses, we can make inferences with the probability intervals $(p_2 - p_0) \pm t_p \text{se}(p_2 - p_0)$. Second, the formulas for standard errors of complex samples can be applied directly to the mean $(p_2 - p_0)$. Third, the logical structure of this statistic $(p_2 - p_0)$ can be seen more clearly in its application to several distinct problems.

Correlated proportions originate usually in data from complex surveys, and the computations of variance should be appropriate to the sample design. The variance formulas for stratified complex samples can be adopted, but the direct formula has eight terms (Kish 1965, 12.10.3). Instead, it is convenient to translate the problem into a trichotomous variable, with values of $-1, 0, +1$ as in design 2 of Section 2; and the computations of Section 4 used that translation.

Then comparisons between variables and between samples can be facilitated by recourse to the design effects:

$$\text{deft}^2(p_2 - p_0) = \frac{\text{computed variance of } (p_2 - p_0)}{[p_2 + p_0 - (p_2 - p_0)^2] / n}.$$

A few words are needed about limitations on the use of *deft* as a tool for robust approximations. They serve well for clustered and multi-stage samples using ultimate clusters (primary selections) for computing sampling errors. However, we avoided the problem of weighted samples, because their treatment would be too specific and perhaps too complex. Weighting for nonresponse would

not be important for the ratio of *deft* $(p_i - p_j)$ to *deft* (p_i) . However weights for gross inequalities of selection probabilities need specific treatments. Nevertheless, inference and experience indicate that *deft* values are less affected by weights than are the variances and means themselves. Furthermore we conjecture that the relations we found between the values of *deft* $(p_i - p_j)$ and *deft* (p_i) will hold also for weighted data, if these are not extreme or pathological.

An approximate but dependable relation of *deft* $(p_i - p_j)$ to *deft* (p_i) and *deft* (p_j) would be useful to allow inferences from the latter, which are routinely and easily computed, to the former that are not. Several alternative conjectures may seem reasonable, and none can be mathematically derived, nor excluded.

1. *Deft* $(p_i - p_j) = 1$ if no design effect was assumed implicitly in the five publications referenced in Section 1.
2. *Deft* $(p_i) > \text{deft}(p_i - p_j) > 1$ denotes persisting but lower effects than for the *deft* (p_i) for proportions. This happens for "crossclasses" and their comparisons (Kish 1987, 7.1). This also seemed reasonable to several experienced statisticians we polled.
3. *Deft* $(p_i - p_j) = [\text{deft}(p_i) + \text{deft}(p_j)] / 2$ is what we actually found to be a good approximation for all of our data, from different populations and designs. This conjecture seems reasonable, because design effects due to clustering for individual p_i can apply similarly to the variable created from the difference $(p_i - p_j)$ of two of them.
4. Inconsistent results would have been possible, but annoying by preventing inference.

4. EMPIRICAL RESULTS FOR *Deft* $(P_i - P_j)$

Without strong theoretical or mathematical basis for favoring any of the four alternative conjectures, empirical results about *deft* $(p_i - p_j)$ become essential, linking these to the computed values for *deft* (p_i) . These resemble our more familiar conjectures about *deft* $(p_i) = \sqrt{1 + \text{roh}[\bar{b} - 1]}$; their value depends on several factors that affect *roh*, the coefficient of intraclass correlation, in addition to the average cluster size \bar{b} (Kish 1965, 5.4, 8.2). The values of *deft* (p_i) vary greatly between surveys, also between variables for the same survey (Kish, Groves and Krotki 1976; Verma, Scott and O'Muircheartaigh 1980; Verma and Lê 1995). However, survey statisticians gain knowledge from empirical investigations of sampling errors from diverse surveys, which also permit relating the *deft* values of complex statistics to the simpler *deft* (p_i) (Kish L. 1995; Rao and Wu 1985; Rao and Scott 1987). Similarly, to learn about the relation of *deft* $(p_i - p_j)$ to *deft* (p_i) we have here empirical results from many variables and from many surveys.

In this first essay into this field we present data from fourteen surveys, which represent a great variety of situations. Eleven surveys presented as 5 sets of results (Figures 1 and 2 and Tables 1-3) deal with paired differences of categories from single surveys (Type A). Three sets of results (Tables 1-3) come from social surveys, followed by two sets (Figures 1 and 2) from the Demographic and Health Surveys on population data. Finally three other sets, each dealing with two waves of data, each based on two reinterviews with the same respondents (Tables 4, 5 and 6), represent type B designs of comparisons.

Tables:

1. The National Election Study of 1986 of the Institute for Social Research of the University of Michigan, $n = 2,135$.
2. The National Education Longitudinal Study (NELS) of 1988, the National Opinion Research Center of the University of Chicago, $n = 24,355$.
3. The National Longitudinal Study of Labor Market Experience of Youth, conducted by the National Opinion Research Center of the University of Chicago, $n = 5,857$.
4. National Election Studies Panels 1990 and 1992, Survey Research Center, Institute for Social Research, Ann Arbor, MI 48106.
5. Panel Study of Income Dynamics 1983 and 1987, Survey Research Center.
6. Americans' Changing Lives 1986 and 1989, Survey Research Center.

Figures:

1. Demographic and Health Surveys of Morocco, Niger, and Colombia, MACRO International.
2. Population Census of Indonesia, Rural Java strata (unpublished data).

We note the following important, useful, EMPIRICAL results.

- 1) First and foremost: The design effects $\text{deft}(p_i - p_j)$ for the differences are usually NO LESS than the $\text{deft}(p_i)$ for the proportions themselves, and $\text{deft}(p_i - p_j) \approx 0.5 [\text{deft}(p_i) + \text{deft}(p_j)]$ approximately in all cases. They vary together, along with the considerable variation for deft values between variables, and also with the lesser variation between pairs of categories for the same variables. Researchers who neglect deft commit the usual under-statement of sampling errors for statistics from clustered surveys. This observation is not only interesting but also a useful model for inference, because the other three sources of variation – across variables, categories within variables, and sampling errors of individual statistics – are all greater.
- 2) We can find these results in all the 14 sets of survey data in the tables and graphs, and we can illustrate them now

with Table 1. Note that defts vary from essentially 1.00 for variable D (problems in country) to as high as 2.32 in variable A (religion) which implies $\text{deft}^2 = 2.32^2 = 5.38$. That our empirical rule (1) holds over the range is reassuring. Such variation between variables in the same sample are common and should force us to abandon the practice of using a common average for all defts of a sample (Verma and Lê 1995; Kish 1995).

Furthermore, we emphasize here the great variation in deft values for the five categories of the same variable from 1.21 to 2.32 (No. 3 for “fundamental” protestants). It follows that $\text{deft}(p_i - p_j)$ is large only when i or j is category 3 for this variable. These variations among the defts for categories of the same variable mean that they should be computed for all categories rather than for only a single “representative” category. These large possible variations between categories of the same variable are an important new finding in our results, that seems to have escaped notice before.

- 3) There are also sampling errors in the computed values of the defts . Only statisticians who have computed many sampling errors and design effects seem to get the “feel” for how great these can be. They may be mostly responsible for the few cases where $\text{deft}(p_i - p_j)$ fails to fall between $\text{deft}(p_i)$ and $\text{deft}(p_j)$ and either $\text{deft}(p_i) < \text{deft}(p_i - p_j) > \text{deft}(p_j)$ or $\text{deft}(p_i) > \text{deft}(p_i - p_j) < \text{deft}(p_j)$. Incidentally, these cases also show that our results are not mathematical consequences, but empirically based.

The empirical results presented in Figures 1 and 2 further confirm the findings already presented in Tables 1, 2, and 3. Here also we see that: 1) $\text{deft}(p_i - p_j) \approx [\text{deft}(p_i) + \text{deft}(p_j)]/2$ approximately, along the 45° line; that 2) those equalities hold along a wide range of designs effects; and that 3) the variation between variables is large indeed. This large variation is particularly evident for rural Indonesia, with deft values over 4, hence deft^2 values over 16. These large clustering effects are due to the large cluster sizes: with $\bar{b} = 133$ and 137, the values of $\text{roh} = 0.12$ are enough for large defts . Note that these empirical results come both from very diverse populations and diverse variables; different from each other and from the data of Tables 1, 2, and 3. Figure 1 has data from 3 countries (Morocco, Niger and Colombia) hence 6 populations, because the urban and rural defts are quite different. Figure 2 shows results for males and females who are quite distinct populations for the occupational variables, though less so for the educational classes.

The empirical data in the tables of studies 4, 5, and 6 were awaited with doubt and anxiety. True that the preceding five sets resulted in similar conclusions, although they dealt with eleven different populations and scores and variables. But studies 1 to 5 all dealt with pairs of categories from polytomies, designs 1 and 2 of Type A. But now we

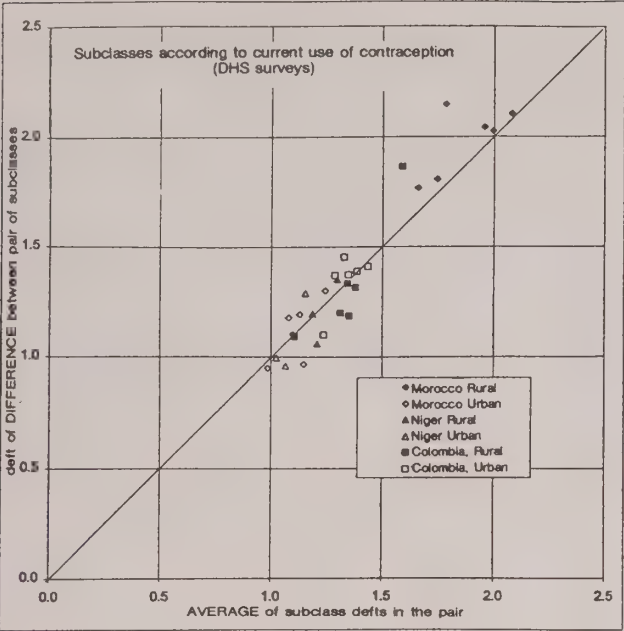


Figure 1. Comparison of $\text{deft}(p_i - p_j)$ to the average of $\text{deft}(p_i), \text{deft}(p_j)$ for categories by current use of contraception*. Illustration of six populations from Demographic and Health Surveys.

- * 1 = not using any method of contraception
- 2 = using only traditional method
- 3 = using a modern 'reversible' method
- 4 = sterilised.

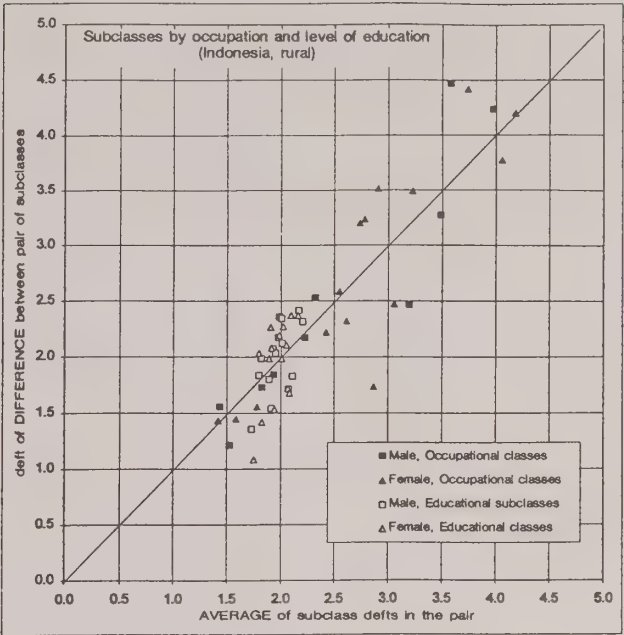


Figure 2. Comparison of $\text{deft}(p_i - p_j)$ to the average of $\text{deft}(p_i), \text{deft}(p_j)$ for categories by occupation and level of education by sex. Illustration from a population census.

Table 1
The National Election Study of 1986 of the I.S.R. of the University of Michigan ($n = 2,135$)

Categories					Categories				
Defts for					Defts for				
$i - j$	P_i	P_j	Average	$(P_i - P_j)$	$i - j$	P_i	P_j	Average	$(P_i - P_j)$
A. Religion					B. Abortions Beliefs				
1-2	1.21	1.42	1.32	1.10	1-2	1.27	.97	1.12	.97
1-3	1.21	2.32	1.77	2.02	1-3	1.27	1.28	1.28	1.32
1-4	1.21	1.50	1.36	1.18	1-4	1.27	1.31	1.29	1.36
1-5	1.21	1.18	1.19	1.17	2-3	.97	1.28	1.12	1.08
2-3	1.42	2.32	1.87	1.93	2-4	.97	1.31	1.14	1.16
2-4	1.42	1.50	1.46	1.57	3-4	1.28	1.31	1.30	1.32
2-5	1.42	1.18	1.30	1.27	Mean	1.17	1.24	1.21	1.20
3-4	2.32	1.50	1.91	2.03	D. Problems in Country				
3-5	2.32	1.18	1.75	2.04	1-2	1.07	.94	1.00	.98
4-5	1.50	1.18	1.34	1.19	1-3	1.07	1.04	1.05	1.09
Mean	1.56	1.53	1.54	1.55	1-4	1.07	.93	1.00	1.12
C. Support Reagan					2-3	.94	1.04	.99	1.01
1-2	1.32	1.10	1.21	1.07	2-4	.94	.93	.93	.85
1-3	1.32	.86	1.09	1.26	3-4	1.04	.93	.98	.82
1-4	1.32	1.48	1.40	1.50	Mean	1.02	.97	.99	.98
2-3	1.10	.86	.98	.96					
2-4	1.10	1.48	1.29	1.38					
3-4	.86	1.48	1.17	1.09					
Mean	1.17	1.21	1.19	1.21					
Overall mean	1.23	1.24	1.23	1.24					

Table 2

The National Education Longitudinal Study (NELS) of 1988, the National Opinion Research Center of the University of Chicago, ($n = 24,355$)

Categories $i - j$	Defts for				Categories $i - j$	Defts for			
	P_i	P_j	Average	$(P_i - P_j)$		P_i	P_j	Average	$(P_i - P_j)$
A. Education Status					B. Classes are boring				
1-2	1.38	1.22	1.30	1.11	1-2	.99	1.11	1.05	1.04
1-3	1.38	1.14	1.26	1.16	1-3	.99	1.12	1.06	1.07
1-4	1.38	1.19	1.29	1.30	2-3	1.11	1.12	1.12	1.13
1-5	1.38	1.42	1.40	1.54	Mean	1.03	1.12	1.08	1.08
2-3	1.22	1.14	1.18	1.11	C. Freedom to pursue interest				
2-4	1.22	1.19	1.21	1.24	1-2	1.28	1.10	1.19	1.21
2-5	1.22	1.42	1.32	1.45	1-3	1.28	1.08	1.18	1.28
3-4	1.14	1.19	1.17	1.18	2-3	1.10	1.08	1.09	.97
3-5	1.14	1.42	1.28	1.37	Mean	1.22	1.09	1.15	1.15
4-5	1.19	1.42	1.31	1.20	D. School offers good jobs				
Mean	1.27	1.28	1.27	1.25	1-2	1.24	1.07	1.16	1.17
E. Religion					1-3	1.24	1.11	1.18	1.24
1-2	2.48	2.83	2.65	2.74	2-3	1.07	1.11	1.09	1.01
1-3	2.48	2.02	2.25	2.09	Mean	1.18	1.10	1.14	1.14
2-3	2.83	2.02	2.42	2.59	F. Dad education				
Mean	2.60	2.29	2.44	2.47	1-2	1.61	1.76	1.69	1.83
I. Feel good about self					1-3	1.61	1.68	1.65	1.65
1-2	1.42	1.28	1.35	1.37	2-3	1.76	1.68	1.72	2.48
					Mean	1.65	1.71	1.69	1.99
Overall mean						1.48	1.41	1.45	1.49

Table 3

The National Longitudinal of Labor Market Experience of Youth, Conducted by the National Opinion Research Center of the University of Chicago, ($n = 5,857$)

Categories	Defts for				Categories	Defts for			
$i - j$	P_i	P_j	Average	$(P_i - P_j)$	$i - j$	P_i	P_j	Average	$(P_i - P_j)$
A. Chance is important in my life					B. Something stops me				
1-2	1.26	1.20	1.23	1.06	1-2	1.07	1.22	1.14	1.04
1-3	1.26	1.18	1.22	1.30	1-3	1.07	1.12	1.10	1.14
1-4	1.26	1.16	1.21	1.28	1-4	1.07	1.09	1.08	1.09
2-3	1.20	1.18	1.19	1.22	2-3	1.22	1.12	1.17	1.28
2-4	1.20	1.16	1.18	1.25	2-4	1.22	1.09	1.16	1.14
3-4	1.18	1.16	1.17	1.05	3-4	1.12	1.09	1.11	1.07
Mean	1.23	1.17	1.20	1.19	Mean	1.13	1.12	1.13	1.13
C. Have control of my life					D. I am as worthy as others				
1-2	1.13	1.06	1.10	1.09	1-2	1.17	1.13	1.15	1.16
1-3	1.13	1.10	1.12	1.13	1-3	1.17	1.07	1.12	1.16
2-3	1.06	1.10	1.08	1.07	2-3	1.13	1.07	1.10	1.08
Mean	1.11	1.09	1.10	1.10	Mean	1.16	1.09	1.12	1.13
E. Plans hardly work out					F. I am satisfied				
1-2	1.19	1.07	1.13	1.12	1-2	1.19	1.12	1.16	1.16
1-3	1.19	1.13	1.16	1.20	1-3	1.19	1.13	1.16	1.20
2-3	1.07	1.13	1.10	1.08	2-3	1.12	1.13	1.13	1.09
Mean	1.15	1.11	1.13	1.13	Mean	1.17	1.13	1.15	1.15
I. Mother's work									
1-2	1.49	1.36	1.43	1.41					
1-3	1.49	1.52	1.51	1.53					
2-3	1.36	1.52	1.44	1.44					
Mean	1.45	1.47	1.46	1.47					
Overall mean	1.20	1.17	1.18	1.19					

Table 4

National Election Studies Panels 1990 and 1992,
Survey Research Center, Institute for Social Research,
Ann Arbor

Categories before/after (90/92)	Defts for			
	P_i	P_j	Average	$(P_i - P_j)$
Strongly approve Bush	1.14	.93	1.04	1.02
Approve Bush foreign policy	.92	1.05	.99	1.00
Strongly disapprove Bush foreign policy	1.23	1.24	1.24	1.32
Approve Bush economy	.97	.94	.96	.96
Strongly approve Bush economy	1.14	1.04	1.09	1.10
Approve Bush	1.00	1.00	1.00	1.00
Strongly disapprove Bush	1.16	1.10	1.13	1.12
Watch campaign on TV	.89	1.55	1.22	1.40
<i>Mean</i>	<i>1.06</i>	<i>1.11</i>	<i>1.08</i>	<i>1.11</i>

Table 5

Panel Study of Income Dynamics, 1983 and 1987,
Survey Research Center, Ann Arbor

Categories* before/after (83/87)	Defts for			
	P_i	P_j	Average	$(P_i - P_j)$
Live in South	1.22	1.23	1.23	1.11
Age of head of family	1.28	1.33	1.31	1.37
Family size	1.29	1.43	1.36	1.47
Number of children in family	1.23	1.43	1.33	1.49
Work hours of head	1.12	.84	.98	1.03
Age of youngest child	.93	.91	.92	.87
<i>Mean</i>	<i>1.18</i>	<i>1.20</i>	<i>1.19</i>	<i>1.22</i>

* All variables are categorized in two categories.

sought data for Type B comparisons from panel surveys, so that we could investigate the conjectures for the test/retest and before/after experimental designs. Mathematically these can be easily shown to resemble polytomies (*i.e.*, tetratomies), but from that to the empirical values of design effects leads through a "black box." Hence these empirical values are so much more valuable and remarkable. Here we found considerable design effects for Chi square tests for analytical comparisons.

5. PRESENT FINDINGS IN THE CONTEXT OF RELATED RESEARCH

A great deal of empirical information is available from previous work by the authors and by others on design effects for the total sample, for subclasses, and for differences, for diverse variables and designs. It would be useful to put the present findings in the context of that work.

It has been found that nature of the survey variables being estimated is a major (often the main) determinant of the magnitude of the design effects: vastly differing defts can occur for different types of variables even with the same samples or with similar designs. For this reason we have always recommended that defts be computed for many different variables, while it is generally less important to compute them for many different subclasses, especially for different categories of subclasses defined in terms of the same characteristic.

The present findings illustrate that defts can differ greatly also among different categories of the same survey variable, estimated with the total sample as the common base. Therefore each individual category and each difference between pairs of categories, even when defined in

Table 6

Americans' Changing Lives, 1986 and 1989, Survey Research Center, Ann Arbor

Categories before/after	Defts for				Categories before/after	Defts for			
	P_i	P_j	Average	$(P_i - P_j)$		P_i	P_j	Average	$(P_i - P_j)$
A. Get together with friends					B. How often do you exercise				
Once a week	1.30	1.26	1.28	1.28	Often	1.51	1.67	1.59	1.26
2-3 a month	.88	1.00	.94	1.02	Never	1.62	1.97	1.80	1.41
<i>Mean</i>	<i>1.09</i>	<i>1.13</i>	<i>1.11</i>	<i>1.15</i>	<i>Mean</i>	<i>1.56</i>	<i>1.82</i>	<i>1.70</i>	<i>1.34</i>
C. How Satisfy Are You					D. How do you like your home				
Very satisfy	1.28	1.21	1.25	1.33	Very much	1.24	.90	1.07	.91
Not satisfy	1.04	1.16	1.10	1.00	Not much	1.33	.98	1.16	1.12
<i>Mean</i>	<i>1.16</i>	<i>1.19</i>	<i>1.18</i>	<i>1.17</i>	<i>Mean</i>	<i>1.29</i>	<i>.94</i>	<i>1.12</i>	<i>1.02</i>
E. How often work in garden					F. I have a positive attitude				
Often	1.40	1.16	1.28	1.19	Agree	1.10	1.33	1.22	1.19
Rarely	.91	1.11	1.01	1.18	Disagree	1.05	1.28	1.17	1.21
Never	1.66	1.17	1.42	1.26	<i>Mean</i>	<i>1.08</i>	<i>1.31</i>	<i>1.20</i>	<i>1.20</i>
<i>Mean</i>	<i>1.32</i>	<i>1.15</i>	<i>1.24</i>	<i>1.21</i>					
<i>Overall mean</i>	<i>1.25</i>	<i>1.26</i>	<i>1.26</i>	<i>1.18</i>					

terms of the same survey variable, needs to be regarded, in a sense, as a separate variable in its own right for the purpose of computing and analyzing design effects.

As to the relationship between defts for subclasses and subclass differences, previous research has mostly dealt with the following situation. With the total sample n partitioned into subclasses i of size $n_i = p_i \cdot n$, $\text{deft}(r_i)$ values for statistics r_i (such as a proportion m_i/n_i , mean $\sum y_i/n_i$, ratio $\sum y_i/\sum x_i$), estimated over subclass elements n_i as the base, are related to $\text{deft}(r)$ for the same variable estimated with the total sample as the base. Similarly, $\text{deft}(r_i - r_j)$ for subclass differences are related to $\text{deft}(r_i)$, $\text{deft}(r_j)$ based on individual subclasses and to $\text{deft}(r)$ based on the total sample. Numerous computations confirm these relationships to be in accord with our conjecture (2) of section 3:

$$\text{deft}(r) > \text{deft}(r_i); \text{ and } \text{deft}(r_i) > \text{deft}(r_i - r_j) > 1.$$

These effects of covariances on design effects of clustered samples are essentially empirical (even sociological in a broad sense); and they must be so verified.

Similarly with our newly discovered relationship for $(p_i - p_j)$ for two categories, which are so different from the above. The relations $\text{deft}(p_i - p_j) \cong [\text{deft}(p_i) + \text{deft}(p_j)]/2$ are also empirical and approximate and they must be verified over and over again. But they seem to be widely applicable in our data, and clearly better than the other assumptions, such as $\text{deft}(p_i - p_j) = 1$ that have been often assumed until now.

ACKNOWLEDGEMENTS

The authors wish to thank the editor and referees whose suggestions made this paper both shorter and better.

REFERENCES

- COCHRAN, W.G. (1950). The comparison of percentages in matched samples. *Biometrika*, 37, 256-66.
- DEMING, W.E. (1953). On the distinction between enumerative and analytic studies. *Journal of the American Statistical Association*, 48, 244-45.
- KISH, L. (1965). *Survey Sampling*. New York: John Wiley and Sons.
- KISH, L. (1987). *Statistical Research Design*. New York: John Wiley and Sons.
- KISH, L. (1995). Methods for design effects. *Journal of Official Statistics*, 11, 55-77.
- KISH, L., and FRANKEL, M.R. (1974). Inference from complex samples. *Journal of the Royal Statistical Society*, (B), 36, 1-37.
- KISH, L., GROVES, R.M., and KROTKI, K. (1976). *Sampling Errors for Fertility Surveys*. Occasional Paper No. 17, *World Fertility Surveys*. International Statistical Institute: The Hague.
- LÊ, T., and VERMA, V. (1995). *Sample Designs and Sampling Errors for the DHS*. Calverton MD: MACRO International.
- McNEMAR, Q. (1949). *Psychological Statistics*. New York: John Wiley and Sons.
- MOSTELLER, F. (1952). Some statistical problems in measuring the subjective responses to drugs. *Biometrika*, 8, 220-226.
- RAO, J.N.K., and SCOTT, A.J. (1987). On simple adjustments to Chi-square tests with sample survey data. *Annals of Statistics*, 15, 385-397.
- RAO, J.N.K., and WU, C.F.S. (1985). Inference from stratified samples: Second-order analysis of three methods for nonlinear statistics. *Journal of the American Statistical Association*, 80, 620-630.
- SCOTT, A.J., and HOLT, D. (1982) The effect of two-stage sampling on ordinary least squares methods. *Journal of the American Statistical Association*, 77, 848-54.
- SCOTT, A.J., and SEBER, G.A.F. (1983). Difference of proportions from the same survey. *The American Statistician*, 37, 319-20.
- SKINNER, C.J., HOLT, D., and SMITH, T.M.F. (1989). *Analysis of Complex Surveys*. New York: John Wiley and Sons.
- VERMA, V., and LÊ, T. (1995). Sampling errors for the DHS survey. 50th Session of the International Statistical Institute, Beijing.
- VERMA, V., SCOTT, C., and O'MUIRCHARTAIGH, C. (1980). Sample designs and sampling errors for the World Fertility Surveys. *Journal of the Royal Statistical Society (A)*, 143, 431-473.
- WILD, C.J., and SEBER, G.A.F. (1993). Comparing two proportions for the same survey. *The American Statistician*, 47, 178-181.

Alternative Adjustments Where There Are Several Levels of Auxiliary Information

F. DUPONT¹

ABSTRACT

Regression estimation and its generalization, calibration estimation, introduced by Deville and Särndal in 1993, serves to reduce *a posteriori* the variance of the estimators through the use of auxiliary information. In sample surveys, there is often useable supplementary information that is distributed according to a complex schema, especially where the sampling is realized in several phases. An adaptation of regression estimation was proposed along with its variants in the framework of two-phase sampling by Särndal and Swensson in 1987. This article seeks to examine alternative estimation strategies according to two alternative configurations for auxiliary information. It will do so by linking the two possible approaches to the problem: use of a regression model and calibration estimation.

KEY WORDS: Auxiliary information; Regression estimator; Calibration estimator; Two-phase sampling.

1. INTRODUCTION

Using the regression estimator studied by Fuller (1975), Cassel, Särndal and Wretman (1976), Särndal (1980), Gourieroux (1981), Isaki and Fuller (1982), and Wright (1983), it is possible to improve *a posteriori* – that is, after the sampling has been completed – the estimate of a total of a variable of interest on the basis of auxiliary variables x_1, \dots, x_k for which additional information is available. The variance in relation to the Horwitz-Thompson estimator is reduced by using the regression estimator, provided that one knows the true value of the target population totals of the auxiliary variables, which will constitute the additional information referred to as auxiliary information. Deville and Särndal in 1992 proposed a class of estimators derived from a reweighting approach that addresses the same issue of variance reduction: calibration estimators. By calibrating sampling weights it is possible to incorporate *a posteriori* auxiliary information of the type totals X_1, \dots, X_k of k variables x_1, \dots, x_k into the estimate made on the basis of the new weightings and thus to improve the estimate. This approach generalizes regression estimation, which is one of the elements of the class.

However, in surveys based on sampling, there is often usable additional information that is distributed according to a more complex schema than what has been described above, especially when the sampling is carried out in several phases. This article looks at different strategies for using this complex auxiliary information in the framework of two-phase sampling, with the possibility of generalizing to more than two phases.

When the sampling plan entails two phases, the auxiliary information consists of information known for the entire population, but also of information known for the

sample resulting from the first sampling phase. These two bodies of information may concern different variables.

In their 1987 article, Särndal and Swensson propose an estimator that uses all the auxiliary information available for a two-phase sampling, with different auxiliary information for the total population and the population obtained from the first-phase sampling. This is an estimator adapting the principle of the regression estimator when the information known for the individuals obtained from the first-phase sampling is considered to be substitutable for the aggregated information and to be of better quality than the information available for the target population as a whole, for purposes of estimating the variable of interest. However, in practice it often happens that these two bodies of information are complementary rather than substitutable. We have thus sought in this study to develop the regression estimate in a context in which the bodies of auxiliary information are complementary.

Furthermore, insofar as calibration estimation generalizes regression estimation when the auxiliary information is at only one level, we have sought to adapt calibration estimation to this context. We review the various calibration strategies in order to propose the most suitable ones, seeking to relate them to generalizations of regression estimation that are possible in this context.

We show (Section 2) that the joint use of two different bodies of auxiliary information leads to two regression models and three associated decompositions of the variable of interest. The regression model assisted approach (RMAA) thus enables us to derive 3 alternative estimators.

In turn, the calibration approach (CA) (Section 3) enables us to derive 4 estimators. Each of these estimators may be related to (associated with) the three estimators derived from the regression model approach.

¹ F. Dupont, Unité Méthodes Statistiques, Institut National de la Statistique et des Études Économiques, (INSEE), 18 Blvd. Adolphe Pinard, 75675 Paris Cedex.

Thus (Section 4), the two approaches may be linked together and result in three classes of estimators, each associated with a decomposition of the variable of interest. The estimators of a given class have the same asymptotic variance.

When strategies are evaluated on the basis of the sampling plan alone, our choice is directed toward the third class of estimators, which is superior to the other two from the standpoint of variance.

When strategies are evaluated on the basis of a modelling of the variable of interest, the preferable class of estimators is the one associated with the modelling adopted.

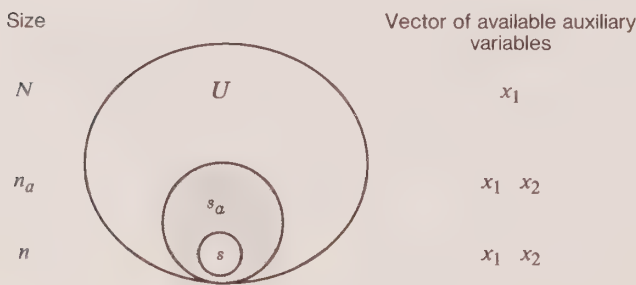
In a situation in which we wish to adjust a survey, and in which we wish simultaneously to correct the biases that would result from the use of gross weightings and to reduce the variance, the findings must be adapted: the changes introduced in the weightings to correct the biases are greater than the corrections for variance reduction. Hence the variables will be incorporated into the calibration once it appears that they are affecting the probability of selection and thus participating in the creation of the bias.

When the auxiliary variables are qualitative, the choice between *a priori* and *a posteriori* use of the auxiliary information – that is, between its use at the sampling stage and at the adjustment stage – still rests on the distinction between the two modellings of the variable of interest.

These findings may be extended to samplings of more than two phases.

2. NOTATIONS

The framework is that of a two-phase sampling. Assume that auxiliary information is available at two different levels: the target population and the population obtained from the first-phase sampling. The situation may be diagrammed as follows:



where U represents the target population for which the values of the vector of variable x_1 are known or, failing that, the total $X_1 = \sum_{i \in U} x_{i1}$. s_a represents an intermediate level of sampling for which the values of the vectors of k_1 variables x_1 and k_2 variables x_2 are known for all individuals. We denote as π_{ia} the probability of selection

from the sample associated with the first phase of the sampling. s represents the final sample for which are available the values of the variable y , the total of which we are trying to estimate, as well as the values of the vectors of the auxiliary variables x_1 and x_2 . This is denoted as $\pi_i = P(i | s_a)$.

We hope to make optimum use of all this auxiliary information in order to improve the estimates that will be made on the basis of the data gathered from the sample that results from the second sampling phase s .

An obvious first idea is to try to generalize the regression estimator in this context.

3. REGRESSION ESTIMATION APPROACH

3.1 The Information Contained in x_1 is Considered to be Substitutable for the Information Contained in x_2 for Estimating y and to be of Lesser Quality

In their work, Särndal, Swensson and Wretman propose the following regression estimator for estimating the total of y :

$$\hat{Y}_1 = \sum_{i \in s} \frac{y_i}{\pi_i \pi_{ai}} + \left(\sum_{i \in s_a} \frac{x'_{i2} \hat{b}_2}{\pi_i} - \sum_{i \in s} \frac{x'_{i2} \hat{b}_2}{\pi_i \pi_{ai}} \right) + \left(\sum_{i \in U} x'_{i1} \hat{b}_1 - \sum_{i \in s_a} \frac{x'_{i1} \hat{b}_1}{\pi_{ai}} \right)$$

where the second term is the correction for poor estimation on s_a and the third is the correction for poor estimation on s .

The estimation can also be written:

$$\hat{Y}_1 = \sum_{i \in U} x'_{i1} \hat{b}_1 + \sum_{i \in s_a} \frac{(x'_{i1} \hat{b}_1 - x'_{i2} \hat{b}_2)}{\pi_{ai}} + \sum_{i \in s} \frac{(y_i - x'_{i2} \hat{b}_2)}{\pi_i \pi_{ai}}$$

where the second term is the correction for poor approximation of y_i by $x'_{i1} \hat{b}_1$ and the third is the correction for poor approximation of y_i by $x'_{i2} \hat{b}_2$;

$$\text{with } \hat{b}_1 = \left(\sum_{i \in s_a} \frac{x_{i1} x'_{i1}}{\pi_{ai}} \right)^{-1} \left(\sum_{i \in s} \frac{x_{i1} y_i}{\pi_i \pi_{ai}} \right)$$

$$\text{and } \hat{b}_2 = \left(\sum_{i \in s} \frac{x_{i2} x'_{i2}}{\pi_i \pi_{ai}} \right)^{-1} \left(\sum_{i \in s} \frac{x_{i2} y_i}{\pi_i \pi_{ai}} \right).$$

The underlying idea is that we have two concurrent models for y , namely:

- (1) $y_i = x'_{i1}b_1 + u_{i1}$ with $E(u_{i1}) = 0$ and $V(u_{i1}) = \sigma_1^2$ and
 (2) $y_i = x'_{i2}b_2 + u_{i2}$ with $E(u_{i2}) = 0$ and $V(u_{i2}) = \sigma_2^2$

the second of which we believe is *a priori* better for predicting the value of y_i . Thus in this model-based perspective, x_1 functions as a proxy of x_2 . A situation of this type corresponds, for example, to a case in which x_2 represents the update – that is, the update to the date of the survey – of the variable retrieved from the x_1 sampling frame. In other words, if x_2 were available at the level of the entire population, the estimator used would be

$$\sum_{i \in U} x'_{i2} \hat{b}_2 + \sum_{i \in s} \frac{(y_i - x'_{i2} \hat{b}_2)}{\pi_i \pi_{ai}}.$$

Let us now imagine the case of a two-phase sampling survey of households. Assume that the sampling frame is made up of dwellings for which we have information consisting of dwelling size, denoted as x_1 , which is therefore known for all individuals in the target population. If all the individuals obtained from the first sampling phase are questioned on the composition of the household, denoted as x_2 , in particular on the number of children in the household, the two bodies of information appear to be complementary rather than substitutable for purposes of studying the household budget. This is further reinforced if instead of household composition, the information collected is the age or occupation of the head of household.

In a model-based perspective, the alternative situation, in which the information contained in x_1 is considered complementary to that contained in x_2 for estimating y , thus naturally suggests itself.

3.2 The Information Contained in x_1 is Considered to be Complementary to the Information Contained in x_2 for Estimating y

3.2.1 Decomposition $y_i = x'_{i1}a_1 + x'_{i2}a_2 + u_i$

The underlying model is then:

$$y_i = x'_{i1}a_1 + x'_{i2}a_2 + u_i \text{ with } E(u_i) = 0 \text{ and } V(u_i) = \sigma_1^2.$$

The estimator to be used is then:

$$\hat{Y}_2 = \sum_{i \in U} x'_{i1} \hat{a}_1 + \sum_{i \in s_a} \frac{x'_{i2} \hat{a}_2}{\pi_{ai}} + \sum_{i \in s} \frac{(y_i - x'_{i1} \hat{a}_1 - x'_{i2} \hat{a}_2)}{\pi_i \pi_{ai}}$$

with:

$$\begin{pmatrix} \hat{a}_1 \\ \hat{a}_2 \end{pmatrix} = \begin{pmatrix} \sum_{i \in s} \frac{x_{i1} x'_{i1}}{\pi_{ai} \pi_i} & \sum_{i \in s} \frac{x_{i1} x'_{i2}}{\pi_{ai} \pi_i} \\ \sum_{i \in s} \frac{x_{i2} x'_{i1}}{\pi_{ai} \pi_i} & \sum_{i \in s} \frac{x_{i2} x'_{i2}}{\pi_{ai} \pi_i} \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i \in s} \frac{x_{i1} y_i}{\pi_{ai} \pi_i} \\ \sum_{i \in s} \frac{x_{i2} y_i}{\pi_{ai} \pi_i} \end{pmatrix}.$$

The variable here is broken down into three components $y_i = x'_{i1} \hat{a}_1 + x'_{i2} \hat{a}_2 + \hat{u}_i$. The total of y is thus broken down into three components, each of which is estimated at the highest level, that is, with the greatest precision possible:

- U for $x'_{i1} \hat{a}_1$,
- s_a for $x'_{i2} \hat{a}_2$, and
- s for \hat{u}_i .

3.2.2 Decomposition $y_i = x'_{i1}c_1 + M_{x_1}(x_{i2})'c_2 + u_i$

If we wish to make maximum use of the information contained in x_1 available on U , it is natural to introduce another formulation of the same model $y_i = x'_{i1}a_1 + x'_{i2}a_2 + u_i$ which isolates everything which in y can be taken into account through x_1 . It is written as follows:

$$y_i = x'_{i1}c_1 + M_{x_1}(x_{i2})'c_2 + u_i \text{ with}$$

$$E(u_i) = 0 \text{ and } V(u_i) = \sigma_1^2,$$

where M_{x_1} represents the orthogonal projection, in the metric associated with the weights $1/\pi_{ai}$, on the orthogonal of the vector space generated in s_a (similar to \mathfrak{R}^n) by the group of variables x_1 .

$M_{x_1}(x_{i2})$ is defined by:

$$M_{x_1}(x_{i2}) = x'_{i2} - \left(\sum_{i \in s_a} \frac{x_{i2} x'_{i1}}{\pi_{ai}} \right) \left(\sum_{i \in s_a} \frac{x_{i1} x'_{i1}}{\pi_{ai}} \right)^{-1} x'_{i1}.$$

The associated natural estimator is then:

$$\hat{Y}_3 = \sum_{i \in U} x'_{i1} \hat{c}_1 + \sum_{i \in s_a} \frac{(M_{x_1} x'_{i2} \hat{c}_2)}{\pi_{ai}} + \sum_{i \in s} \frac{(y_i - x'_{i1} \hat{c}_1 - M_{x_1} x'_{i2} \hat{c}_2)}{\pi_i \pi_{ai}},$$

where $\hat{c} = \begin{pmatrix} \hat{c}_1 \\ \hat{c}_2 \end{pmatrix}$ is the regression coefficient $y = x'c_1 + (M_{x_1}x_2)'c_2 + u$ estimated over s with weights $1/\pi_{ai}\pi_i$ (which differs slightly from $\begin{pmatrix} \hat{b}_1 \\ \hat{b}_2 \end{pmatrix}$).

3.3 The Three Estimators Derived from the Model-based Approach

The modelling approach has enabled us to construct 3 estimators that can be rewritten synthetically by introducing

new notations. Throughout what follows, for a vector of a given variable z , the following notation will be used:

$$\hat{\hat{Z}} = \sum_{i \in s} \frac{1}{\pi_{ai} \pi_i} z$$

$$\hat{Z} = \sum_{i \in s_a} \frac{1}{\pi_{ai}} z.$$

With these notations, the three estimators are rewritten as follows:

$$\hat{Y}_1 = [X'_1 \hat{b}_1] + [\hat{X}'_2 \hat{b}_2 - \hat{X}'_1 \hat{b}_1] + [\hat{Y} - \hat{X}'_2 \hat{b}_2]$$

associated with the models:

$$(1) \quad y_i = x'_{i1} b_1 + u_{i1}$$

and

$$(2) \quad y_i = x'_{i2} b_2 + u_{i2}$$

$$\hat{Y}_2 = [X'_1 \hat{a}_1] + [\hat{X}'_2 \hat{a}_2] + [\hat{Y} - \hat{X}'_1 \hat{a}_1 - \hat{X}'_2 \hat{a}_2]$$

associated with the model

$$y_i = x'_{i1} a_1 + x'_{i2} a_2 + u_i$$

$$\hat{Y}_3 = [X'_1 \hat{c}_1] + [M_{x_1} \hat{X}'_2 \hat{c}_2] + [\hat{Y} - \hat{X}'_1 \hat{c}_1 - M_{x_1} \hat{X}'_2 \hat{c}_2]$$

associated with the model

$$y_i = x'_{i1} c_1 + M_{x_1}(x_{i2})' c_2 + u_i.$$

In the same manner as the regression estimator is generalized by calibration estimators, the problem of the use of auxiliary information at several levels may be dealt with through calibration theory, by attempting to construct calibration strategies adapted to the auxiliary information configuration examined in this article.

4. CALIBRATION APPROACH

4.1 Different Strategies Possible

When we try to generalize the calibration estimate proposed in a context in which auxiliary information is present at a single level – that of the entire population – several strategies naturally suggest themselves:

Strategy 1

- calibrate the structure of the 1st-phase sample s_a on that of the total population U in terms of variable x_1 , then,
- calibrate the structure of the 2nd-phase sample s on that of the 1st phase sample s_a in terms of variable x_2 .

Note: For the latter operation, it is better to take account of the preceding calibration in terms of x_1 in order to determine the reference value in the calibration in terms of x_2 on s_a . If the preceding calibration is not taken into account, only the estimates made at the level of s_a will benefit from the improvement made by stage a. A good way to convince oneself of this is to consider the specific extreme case where $x_1 = x_2$.

This strategy corresponds to the following calibration equations:

Stage a:

$$\sum_{i \in s_a} \frac{F(x'_{i1} \beta_1)}{\pi_{ai}} x_{i1} = \sum_{i \in U} x_{i1} = X_1$$

which determines β_1 , then

Stage b:

$$\sum_{i \in s} \frac{F(x'_{i1} \beta_1)}{\pi_{ai} \pi_i} F(x'_{i2} \beta_2) x_{i2} = \sum_{i \in s_a} \frac{F(x'_{i1} \beta_1)}{\pi_{ai}} x_{i2} = \hat{X}_2^*$$

which determines β_2 ,

where F designates, as throughout this article, the function which is used in the calibration and which may be linear, exponential, truncated linear or logit (see Deville, Särndal 1993).

Strategy 2

Calibrate the structure of the 2nd-phase sample s simultaneously in terms of variables x_1 and x_2 , that is,

- on the structure of the total population U as regards x_1
- on the structure of s_a for x_2 .

This second strategy leads us to the following calibration equations:

$$\sum_{i \in s} \frac{F(x_{i1} \alpha_1 + x_{i2} \alpha_2)}{\pi_{ai} \pi_i} x_{i1} = \sum_{i \in U} x_{i1} = X_1, \quad \text{and}$$

$$\sum_{i \in s} \frac{F(x'_{i1} \alpha_1 + x'_{i2} \alpha_2)}{\pi_{ai} \pi_i} x_{i2} = \sum_{i \in s_a} \frac{x_{i2}}{\pi_{ai}} = \hat{X}_2,$$

which determines α_1 and α_2 .

The first strategy offers the advantage of correcting the 1st phase weightings, that is, of incorporating the auxiliary information at the highest level. The second strategy, for its part, makes it possible to correct the weightings that will actually be used in the estimation, and in particular to obtain a perfect estimate of the total of x_1 .

A third strategy may be proposed; it combines the advantages of the above two strategies and would therefore seem preferable to them:

Strategy 3

- a) calibrate the structure of the 1st phase sample s_a on that of the total population U in terms of variable x_1 , then
- b) calibrate the structure of the 2nd phase sample s simultaneously in terms of variables x_1 and x_2 , that is,
 - on the structure of the total population U as regards x_1
 - on the structure of s_a modified by taking account of the preceding calibration for x_2 .

This strategy leads to the following calibration equations:

Stage a:

$$\sum_{i \in s_a} \frac{F(x'_{i1} \beta_1)}{\pi_{ai}} x_{i1} = \sum_{i \in U} x_{i1}$$

which determines β_1 , then

Stage b:

$$\sum_{i \in s} \frac{F(x'_{i1} \gamma_1 + x'_{i2} \gamma_2)}{\pi_{ai} \pi_i} x_{i1} = \sum_{i \in U} x_{i1} = X_1, \quad \text{and}$$

$$\sum_{i \in s} \frac{F(x'_{i1} \gamma_1 + x'_{i2} \gamma_2)}{\pi_{ai} \pi_i} x_{i2} = \sum_{i \in s_a} \frac{F(x'_{i1} \beta_1)}{\pi_{ai}} x_{i2} = \hat{X}_2^*,$$

which determines γ_1 and γ_2 .

Lastly, a fourth strategy may be proposed; it may be seen as a variant of the preceding strategy:

Strategy 4

- a) calibrate the structure of the 1st phase sample s_a on that of the total population U in terms of variable x_1 , then
- b) calibrate the structure of the 2nd phase sample s simultaneously in terms of variables x_1 and x_2 , on the basis of the weights modified by the preceding calibration, that is,
 - on the structure of the total population U as regards x_1
 - on the structure of s_a modified taking account of the preceding calibration for x_2 .

This strategy leads to the following calibration equations:

Stage a:

$$\sum_{i \in s_a} \frac{F(x'_{i1} \beta_1)}{\pi_{ai}} x_{i1} = \sum_{i \in U} x_{i1},$$

which determines β_1 , then

Stage b:

$$\sum_{i \in s} \frac{F(x'_{i1} \beta_1) F(x'_{i1} \delta_1 + x'_{i2} \delta_2)}{\pi_{ai} \pi_i} x_{i1} = \sum_{i \in U} x_{i1} = X_1, \quad \text{and}$$

$$\sum_{i \in s} \frac{F(x'_{i1} \beta_1) F(x'_{i1} \delta_1 + x'_{i2} \delta_2)}{\pi_{ai} \pi_i} x_{i2} =$$

$$\sum_{i \in s_a} \frac{F(x'_{i1} \beta_1)}{\pi_{ai}} x_{i2} = \hat{X}_2^*,$$

which determines δ_1 and δ_2 .

When the calibration function is exponential, it is clear that strategies 3 and 4 coincide.

In this calibration-based approach, the viewpoint adopted is that of reduction of variance based on the characteristics of the sampling plan, without consideration of the model. Two questions then naturally arise:

- Can each of these four strategies be linked to a model-based approach?
- Can these four strategies be compared in terms of variance?

We will first examine the link between the three strategies defined by a calibration approach and the strategies defined by a model-based or regression approach, after which we will focus on calculating the variances of the estimators associated with each of the strategies.

4.2 Link Between the Different Possible Strategies and the Regression Approach

When F is linear, each of the estimators associated with the four strategies may be rewritten simply.

Notations

Throughout the rest of this article we will use the following notations for a vector of any variable z :

$$\hat{Z}^* = \sum_{i \in s} \frac{F(x'_{i1} \beta_1)}{\pi_{ai} \pi_i} z_i \quad \hat{Z}^* = \sum_{i \in s_a} \frac{F(x'_{i1} \beta_1)}{\pi_{ai}} z_i.$$

We will also omit the i indexes in order to lighten the presentation when there is no ambiguity.

Strategy 1

The weightings are of the form

$$w_i^4 = \frac{F(x'_{i1} \beta_1)}{\pi_{ai} \pi_i} F(x'_{i2} \beta_2),$$

the associated estimator \hat{Y}_4 may be rewritten by translating the effect of the second calibration on x_2 :

$$\hat{Y}_4 = \hat{Y}^* + [\hat{X}_2^* - \hat{X}_2] \hat{B}_2 \quad \text{with}$$

$$\hat{B}_2 = \left(\sum_s \frac{F(x_1' \beta_1)}{\pi_a \pi} x_2 x_2' \right)^{-1} \left(\sum_s \frac{F(x_1' \beta_1)}{\pi_a \pi} x_2 y \right),$$

then by translating the effect of the first calibration on x_1 :

$$\hat{Y}_4 = \hat{Y} + [X_1 - \hat{X}_1]' \hat{B}_1 + [\hat{X}_2^* - \hat{X}_2] \hat{B}_2,$$

or:

$$\hat{Y}_4 = [X_1' \hat{B}_1] + [\hat{X}_2^* \hat{B}_2 - \hat{X}_1' \hat{B}_1] + [\hat{Y} - \hat{X}_2^* \hat{B}_2],$$

$$\text{with} \quad \hat{B}_1 = \left(\sum_{s_a} \frac{x_1 x_1'}{\pi_a} \right)^{-1} \left(\sum_s \frac{x_1 y}{\pi_a \pi} \right).$$

Now, \hat{Y}_1 is rewritten:

$$\hat{Y}_1 = [X_1' \hat{b}_1] + [\hat{X}_2^* \hat{b}_2 - \hat{X}_1' \hat{b}_1] + [\hat{Y} - \hat{X}_2^* \hat{b}_2].$$

We thus obtain an estimator similar to the estimator \hat{Y}_1 that is obtained from the model-based approach in cases where the information contained in x_1 is considered to be substitutable for the information contained in x_2 for estimating y and also to be of lesser quality. The differences between \hat{Y}_1 and \hat{Y}_4 concern the following points:

1. \hat{B}_2 is estimated by incorporating the changes from the calibration on x_1 , unlike \hat{b}_2 .
2. The estimate $\hat{B}_1 = \left(\sum_{s_a} \frac{x_1 x_1'}{\pi_a} \right)^{-1} \left(\sum_s \frac{x_1 y}{\pi_a \pi} \right)$ of B_1 , is made in part on s_a , unlike \hat{b}_1 .
3. Lastly, we use the adjusted weights $F(x_1 \beta_1) / \pi_a \pi$ in the sums in x_2 on s and on S_a in \hat{Y}_4 in unlike what was done for \hat{Y}_1 : the estimation on x_2 is improved by the knowledge of x_1 .

Thus the underlying modelling here is indeed: (1) $y_i = x_{i1}' b_1 + u_{i1}$ and (2) $y_i = x_{i2}' b_2 + u_{i2}$, the second of which we think is *a priori* better for predicting the value of y_i .

Strategy 2

We obtain weights

$$w_i^5 = \frac{F(x_{i1}' \alpha_1 + x_{i2}' \alpha_2)}{\pi_{ai} \pi_i},$$

the associated estimator is rewritten as follows:

$$\hat{Y}_5 = [X_1' \hat{a}_1] + [\hat{X}_2' \hat{a}_2] + [\hat{Y} - \hat{X}_1' \hat{a}_1 - \hat{X}_2' \hat{a}_2].$$

We thus obtain exactly the estimator \hat{Y}_2 proposed in the regression model approach in the case in which the information contained in x_1 is considered complementary to the information contained in x_2 for estimating y . The underlying model here is indeed $y_i = x_{i1} a_1 + x_{i2} a_2 + u_i$.

Strategy 3

We obtain weights

$$w_i^6 = \frac{F(x_{i1}' \gamma_1 + x_{i2}' \gamma_2)}{\pi_{ai} \pi_i},$$

the associated estimator is rewritten as:

$$\hat{Y}_6 = \hat{Y} + [X_1 - \hat{X}_1]' \hat{a}_1 + [\hat{X}_2^* - \hat{X}_2]' \hat{a}_2$$

thus:

$$\hat{Y}_6 = [X_1' \hat{a}_1] + [\hat{X}_2^* \hat{a}_2] + [\hat{Y} - \hat{X}_1' \hat{a}_1 - \hat{X}_2^* \hat{a}_2].$$

Now,

$$\hat{X}_2^* = \sum_{s_a} \frac{x_2}{\pi_a} + \left(\sum_{s_a} \frac{x_2 x_1'}{\pi_a} \right)^{-1} \left(\sum_{s_a} \frac{x_1 x_1'}{\pi_a} \right) \left[X - \sum_{s_a} \frac{x_1}{\pi_a} \right].$$

From this it can be deduced by replacing in \hat{Y}_6 that:

$$\hat{Y}_6 = [X_1' \hat{C}_1] + [M_{x_1} \hat{X}_2' \hat{a}_2] + [\hat{Y} - \hat{X}_1' \hat{a}_1 - \hat{X}_2' \hat{a}_2],$$

with

$$\hat{C}_1 = \hat{a}_1 + \left(\sum_{s_a} \frac{x_1 x_1'}{\pi_a} \right)^{-1} \left(\sum_{s_a} \frac{x_1 x_2'}{\pi_a} \right) \hat{a}_2.$$

We thus obtain an estimator that is close to the estimator \hat{Y}_3 proposed in the regression model approach in the case in which the information contained in x_1 is considered complementary to the information contained in x_2 for estimating y . The underlying model here is $y_i = x_{i1}' c_1 + M_{x_1}(x_{i2})' c_2 + u_i$. The differences between \hat{Y}_3 and \hat{Y}_6 concern the estimated coefficients: (\hat{C}_1) differs slightly from (\hat{c}_1) and $[\hat{Y} - \hat{X}_1' \hat{a}_1 - \hat{X}_2' \hat{a}_2]$ differs slightly from $[\hat{Y} - \hat{X}_1' \hat{c}_1 - M_{x_1} \hat{X}_2' \hat{c}_2]$. On the other hand, these quantities are asymptotically equivalent.

Strategy 4

We obtain weights

$$w_i^7 = \frac{F(x_{i1}' \beta_1) F(x_{i1}' \delta_1 + x_{i2}' \delta_2)}{\pi_{ai} \pi_i},$$

the associated estimator is rewritten as follows:

$$\hat{Y}_7 = \hat{Y}^* + [X_1 - \hat{X}_1^*]' \hat{a}_1^* + [\hat{X}_2^* - \hat{X}_2^*]' \hat{a}_2^*.$$

By changing the initial weights in $d_i = F(x_{i1}\beta_1)/\pi_{ai}\pi_i$ we obtain in the same manner:

$$\hat{Y}_7 = [X_1' \hat{a}_1^*] + [\hat{X}_2^*]' \hat{a}_2^* + [\hat{Y}^* - \hat{X}_1^*]' \hat{a}_1^* - \hat{X}_2^*]' \hat{a}_2^*.$$

By replacing \hat{X}_2^* by its expression found above, we obtain:

$$\hat{Y}_7 = [X_1' \hat{C}_1^*] + [M_{x_1} \hat{X}_2^*]' \hat{a}_2^* + [\hat{Y}^* - \hat{X}_1^*]' \hat{a}_1^* - \hat{X}_2^*]' \hat{a}_2^*,$$

with

$$\hat{C}_1^* = \hat{a}_1^* + \left(\sum_{s_a} \frac{x_1 x_1'}{\pi_a} \right)^{-1} \left(\sum_{s_a} \frac{x_1 x_2'}{\pi_a} \right) \hat{a}_2^*.$$

Finally, \hat{Y}_7 and \hat{Y}_6 are asymptotically equivalent.

Say that $w = y - x_1' \hat{a}_1^* - x_2' \hat{a}_2^*$. Then $\hat{Y}_7 = \hat{Y}_6 + [\hat{W}^* - \hat{W}]$. Now, asymptotically $[\hat{W}^* - \hat{W}]$ is an infinitely small negligible before \hat{Y}_6 :

$$[\hat{W}^* - \hat{W}] = \left(\sum_s \frac{w x_1'}{\pi_a \pi} \right) \left(\sum_{s_a} \frac{x_1 x_1'}{\pi_a} \right)^{-1} [X_1 - \hat{X}_1],$$

and

$$\left(\sum_s \frac{w x_1'}{\pi_a \pi} \right) \text{ tends toward zero and } [X_1 - \hat{X}_1] = O\left(\frac{1}{\sqrt{m}}\right).$$

Ultimately we obtain $\hat{Y}_7 \cong \hat{Y}_6$.

In conclusion, when the calibration function is exponential, the estimator \hat{Y}_7 coincides exactly with the preceding. When F is linear, \hat{Y}_7 is close to the preceding and thus still corresponds to the regression model approach in the case in which the information contained in x_1 is considered complementary to the information contained in x_2 for estimating y and in which the decomposition $y_i = x_{i1}' c_1 + M_{x_1}(x_{i2})' c_2 + u_i$ is used.

Conclusion: The Three Classes of Estimators

We have just seen that the four strategies derived from a calibration approach could be associated with regression modelling. We thus obtain three classes of estimators:

$Y_4 \cong Y_1$ associated with the models

$$(1) \quad y_i = x_{i1}' b_1 + u_{i1},$$

and

$$(2) \quad y_i = x_{i2}' b_2 + u_{i2}$$

$\hat{Y}_5 = \hat{Y}_2$ associated with the model

$$y_i = x_{i1}' a_1 + x_{i2}' a_2 + u_i$$

$\hat{Y}_6 \cong \hat{Y}_3$ and $\hat{Y}_7 \cong \hat{Y}_3$ associated with the model

$$y_i = x_{i1}' c_1 + M_{x_1}(x_{i2})' c_2 + u_i.$$

The approximation \cong , which indicates that the estimators are attached to the same regression model, takes on its full meaning when we are interested in calculating the variance of these different estimators, since the estimators that are attached to the same regression model have the same asymptotic variance.

5. ESTIMATION OF VARIANCES

Let us consider the variances of the different estimators $\hat{Y}_1, \dots, \hat{Y}_7$ defined above. AV designates the asymptotic variance of an estimator that is obtained when N, n and m tend toward infinity in a constant relationship.

5.1 Estimator \hat{Y}_1 and \hat{Y}_4 : model

$$y_i = x_{i1}' b_1 + u_{i1} \text{ and (2) } y_i = x_{i2}' b_2 + u_{i2}.$$

• Estimator \hat{Y}_1

The variance of this estimator and its estimate are given in the work of Särndal, Swensson and Wretman (1991). The variance breaks down into two terms that measure the amounts of variance due respectively to the first and the second phase of the sampling.

$$AV(\hat{Y}_1) = \left(\sum_{i,j \in U} \Delta_{ij}^1 \frac{u_{i1} u_{j1}}{\pi_i \pi_j} \right) + \left(E_{s_a} \sum_{i,j \in s_a} \Delta_{ij}^2 \frac{u_{2i} u_{2j}}{\pi_i \pi_{ai} \pi_j \pi_{aj}} \right),$$

$$\text{with: } \Delta_{ij}^2 = \pi_{ij} - \pi_i \pi_j,$$

$$\Delta_{ij}^1 = \pi_{aij} - \pi_{ai} \pi_{aj},$$

$$u_{1i} = y_i - x_{i1}' b_1,$$

$$u_{2i} = y_i - x_{i2}' b_2,$$

$$b_1 = \left(\sum_{i \in U} x_{i1} x_{i1}' \right)^{-1} \left(\sum_{i \in U} x_{i1} y_i \right),$$

$$b_2 = \left(\sum_{i \in U} x_{i2} x_{i2}' \right)^{-1} \left(\sum_{i \in U} x_{i2} y_i \right).$$

Thus the variance estimator also breaks down into two terms that estimate the amounts of variance relating to each of the sampling phases. We find that by construction

of \hat{Y}_1 , x_1 serves to reduce the variance brought about by the first phase and x_2 serves to reduce the variance brought about by the second phase.

$$\hat{V}(\hat{Y}_1) =$$

$$\left(\sum_{i \in S} \sum_{j \in S} \frac{\Delta_{ij}^1}{\pi_{ij} \pi_{aij}} \frac{\hat{u}_{1i} \hat{u}_{1j}}{\pi_i \pi_j} \right) + \left(\sum_{i,j \in S} \frac{\Delta_{ij}^2}{\pi_{aij}} \frac{\hat{u}_{2i} \hat{u}_{2j}}{\pi_i \pi_{ai} \pi_j \pi_{aj}} \right),$$

1st phase

2nd phase

$$\text{with: } \hat{u}_{1i} = y_i - x'_{i1} \hat{b}_1,$$

$$\hat{u}_{2i} = y_i - x'_{i2} \hat{b}_2.$$

Such a decomposition is based on the expression $V(\hat{Y}_1) = V(E[\hat{Y}_1 | s_a]) + E(V[\hat{Y}_1 | s_a])$, which will apply for all the other estimators.

• Estimator \hat{Y}_4

The terms of the development to the first order in $1/\sqrt{m}$ of \hat{Y}_1 and \hat{Y}_4 coincide exactly. We can therefore give a more precise meaning to the expression $\hat{Y}_4 \cong \hat{Y}_1$. We deduce from this that $AV(\hat{Y}_1) = AV(\hat{Y}_4)$. Thus:

$$\hat{V}(\hat{Y}_4) =$$

$$\left(\sum_{i,j \in S} \sum_{j \in S} \frac{\Delta_{ij}^1}{\pi_{ij} \pi_{aij}} \frac{\tilde{u}_{1i} \tilde{u}_{1j}}{\pi_i \pi_j} \right) + \left(\sum_{i,j \in S} \frac{\Delta_{ij}^2}{\pi_{aij}} \frac{\tilde{u}_{2i} \tilde{u}_{2j}}{\pi_i \pi_{ai} \pi_j \pi_{aj}} \right),$$

$$\text{with: } \tilde{u}_{1i} = y_i - x'_{i1} \hat{B}_1,$$

$$\tilde{u}_{2i} = y_i - x'_{i2} \hat{B}_2.$$

5.2 Estimators $\hat{Y}_2 = \hat{Y}_5$: model

$$y_i = x'_{i1} a_1 + x'_{i2} a_2 + u_i$$

It is easy to show (see Dupont 1994) that:

$$AV(\hat{Y}_2) = AV(\hat{Y}_5) \cong$$

$$\left(\sum_{i,j \in U} \Delta_{ij}^1 \frac{v_i v_j}{\pi_i \pi_j} \right) + \left(E_{s_a} \sum_{i,j \in s_a} \Delta_{ij}^2 \frac{u_i u_j}{\pi_i \pi_{ai} \pi_j \pi_{aj}} \right),$$

$$\text{with: } v_i = y_i - x'_{i1} a_1,$$

$$u_i = y_i - x'_{i1} a_1 - x'_{i2} a_2$$

$$a = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} \sum_{i \in U} x_{i1} x'_{i1} & \sum_{i \in U} x_{i1} x'_{i2} \\ \sum_{i \in U} x_{i2} x'_{i1} & \sum_{i \in U} x_{i2} x'_{i2} \end{pmatrix} \begin{pmatrix} \sum_{i \in U} x_{i1} y_i \\ \sum_{i \in U} x_{i2} y_i \end{pmatrix}$$

From this we deduce that:

$$\hat{V}(\hat{Y}_2) = \hat{V}(\hat{Y}_5) =$$

$$\left(\sum_{i,j \in S} \frac{\Delta_{ij}^1}{\pi_{ij} \pi_{aij}} \frac{\hat{v}_i \hat{v}_j}{\pi_i \pi_j} \right) + \left(\sum_{i,j \in S} \frac{\Delta_{ij}^2}{\pi_{aij}} \frac{\hat{u}_i \hat{u}_j}{\pi_i \pi_{ai} \pi_j \pi_{aj}} \right),$$

with:

$$\hat{v}_i = y_i - x'_{i1} \hat{a}_1,$$

$$\hat{u}_i = y_i - x'_{i1} \hat{a}_1 - x'_{i2} \hat{a}_2.$$

In this formulation we find that by construction of $\hat{Y}_2 - \hat{Y}_5$, x_1 reduces the variance brought about by the first phase and x_1 and x_2 are used simultaneously to reduce the variance brought about by the second phase.

5.3 Estimators \hat{Y}_3 , \hat{Y}_6 and \hat{Y}_7 : model

$$y_i = x'_{i1} c_1 + M_{x_1}(x'_{i2})' c_2 + u_i$$

We show that $AV(\hat{Y}_6) = AV(\hat{Y}_7) = AV(\hat{Y}_3)$. Thus,

$$AV(\hat{Y}_3) = AV(\hat{Y}_6) = AV(\hat{Y}_7) \cong$$

$$\left(\sum_{i,j \in U} \Delta_{ij}^1 \frac{u_{1i} u_{1j}}{\pi_i \pi_j} \right) + \left(E_{s_a} \sum_{i,j \in s_a} \Delta_{ij}^2 \frac{u_i u_j}{\pi_i \pi_{ai} \pi_j \pi_{aj}} \right),$$

$$u_{1i} = y_i - x'_{i1} c_1 = y_i - x'_{i1} b_1,$$

$$u_i = y_i - x'_{i1} c_1 - M_{x_1} x'_{i2} c_2 = y_i - x'_{i1} a_1 - x'_{i2} a_2.$$

From this we deduce the three variance estimators, which differ owing to different estimated coefficients:

$$\hat{V}(\hat{Y}_3) =$$

$$\left(\sum_{i,j \in S} \frac{\Delta_{ij}^1}{\pi_{ij} \pi_{aij}} \frac{\hat{u}_{1i} \hat{u}_{1j}}{\pi_i \pi_j} \right) + \left(\sum_{i,j \in S} \frac{\Delta_{ij}^2}{\pi_{aij}} \frac{\hat{u}_i \hat{u}_j}{\pi_i \pi_{ai} \pi_j \pi_{aj}} \right),$$

$$\hat{u}_{1i} = y_i - x'_{i1} \hat{c}_1,$$

$$\hat{u}_i = y_i - x'_{i1} \hat{c}_1 - M_{x_1} x'_{i2} \hat{c}_2,$$

$$\hat{V}(\hat{Y}_6) =$$

$$\left(\sum_{i,j \in S} \frac{\Delta_{ij}^1}{\pi_{ij} \pi_{aij}} \frac{\tilde{u}_{1i} \tilde{u}_{1j}}{\pi_i \pi_j} \right) + \left(\sum_{i,j \in S} \frac{\Delta_{ij}^2}{\pi_{aij}} \frac{\tilde{u}_i \tilde{u}_j}{\pi_i \pi_{ai} \pi_j \pi_{aj}} \right),$$

$$\tilde{u}_{1i} = y_i - x'_{i1} \hat{C}_1,$$

$$\tilde{u}_i = y_i - x'_{i1} \hat{a}_1 - x'_{i2} \hat{a}_2,$$

$$\hat{V}(\hat{Y}_7) =$$

$$\left(\sum_{i,j \in s} \frac{\Delta_{ij}^1}{\pi_{ij} \pi_{aj}} \frac{\tilde{u}_{1i} \tilde{u}_{1j}}{\pi_i \pi_j} \right) + \left(\sum_{i,j \in s} \frac{\Delta_{ij}^2}{\pi_{aj} \pi_i \pi_{ai} \pi_j \pi_{aj}} \right),$$

$$\tilde{u}_{1i} = y_i - x'_{i1} \hat{C}_1^*,$$

$$\tilde{u}_i = y_i - x'_{i1} \hat{a}_1^* - x'_{i2} \hat{a}_2^*,$$

$$\hat{C}_1 = \hat{a}_1^* + \left(\sum_{s_a} \frac{x_1 x'_1}{\pi_a} \right)^{-1} \left(\sum_{s_a} \frac{x_1 x'_2}{\pi_a} \right) \hat{a}_2^*.$$

We find that by construction of \hat{Y}_3 , \hat{Y}_6 and \hat{Y}_7 , x_1 is used to achieve *maximum* reduction of the variance brought about by the first phase and x_2 serves to reduce the variance brought about by the second phase.

6. CHOICE OF ESTIMATORS WHERE THERE IS SELECTION BIAS

In practice, when a survey is adjusted, it is not unusual to want not only to improve the estimation, but also and more especially to correct the biases introduced by uncontrolled selections of individuals, such as nonresponse.

We shall examine the case of a two-phase sampling in which the second phase is equivalent to total nonresponse. The weights π_i of the second-phase sampling are thus unknown. The calibration of s will enable us to estimate these probabilities, while reducing the variance (*cf.* Deville and Dupont 1993). However, asymptotically, the corrections of bias to be made to the weights are greater than the changes to be made in order to improve the estimators. It is therefore the implicit response model that will guide the choice between the different estimators:

The implicit response model for the first class of estimators is $p_i = 1/F(x'_{i2} B_2)$.

The implicit response model for the second and third classes of estimators is $p_i = 1/F(x'_{i2} A_2 + x'_{i1} A_1)$.

- Whatever the response model, an evaluation of the three classes of estimators on the basis of the sampling plan alone still indicates that the third is preferable, since it is appropriate for all the response models.
- If the strategies are evaluated on the basis of regression modelling, we will use the first class of estimators only if the response mechanism is well explained by x_2 , that is, $p_i = 1/F(x'_{i2} B_2)$. Now, we have seen that the modelling associated with the first class of estimators takes on its meaning when the variables x_1 and x_2 are highly correlated. It is therefore fairly probable that in this context, the variable x_2 will be sufficient to explain the response mechanism. Should this not be the case, it will be necessary to turn to the third class of estimators.

The comparison between the three strategies may thus be adapted in a context in which we wish to correct the biases introduced by uncontrolled selections. The conclusions remain largely the same.

According to the same principle, it is of course possible to make comparisons between alternative adjustment strategies in the context of samplings that entail more than two phases and one or more uncontrolled selections.

7. A PRIORI AND A POSTERIORI USE OF AUXILIARY INFORMATION

The calibration estimator enables us to improve the estimate *a posteriori*, by reducing the variance and correcting the bias, as noted above. However, we may want to incorporate the auxiliary information *a priori*, at the sampling stage rather than *a posteriori* at the estimation stage. We then encounter, in a more complex context, the classical opposition between stratification and poststratification, well known in the case of single-phase sampling, when all the auxiliary variables are qualitative.

It is possible to transpose the terms of the choice between using the information *a priori* and *a posteriori*, in the sampling and auxiliary information configuration studied, when the auxiliary variables are qualitative. When the auxiliary variables are qualitative, a calibration corresponds exactly to poststratification.

We saw earlier that in order to determine the proper adjustment procedure, it was necessary to distinguish two possible modellings of the variable of interest, depending on whether the information in x_1 and the information in x_2 were considered substitutable or complementary. Each of these two modellings then led to one or more different adjustment procedures. Similarly, these two modellings arise when it is a matter of identifying the best sampling strategy for incorporating the auxiliary information:

- When the information in x_1 and the information in x_2 are substitutable, the modelling of the variable of interest is as follows:
 - (1) $y_i = x_{i1} b_1 + u_{i1}$ and
 - (2) $y_i = x_{i2} b_2 + u_{i2}$ where the second model is better for predicting the value of y_i .

We have seen that the use of the auxiliary information *a posteriori* leads to calibration strategy No. 1, that is, to the first class of estimators. If we wish to take account of the auxiliary information at the sampling stage, it is natural to propose a sampling stratified on x_1 for the first phase and a sampling stratified on x_2 for the second phase.

However, the parallel between the adjustment procedure and the sampling procedure is not complete: in a calibration, only the marginal information in x_1 can be used.

This results in incomplete poststratification (Särndal and Deville 1992). On the other hand, in the sampling procedure proposed as an *a priori* alternative, we are obliged to use all the cross-tabulations of the x_1 variables. The *a priori* equivalent of a calibration would accordingly be a sampling balanced on the margins of the vector of variables x_1 .

- When the information contained in x_1 and the information contained in x_2 are complementary, the modelling of the variable of interest is $y_i = x_{i1}b_1 + x_{i2}b_2 + u_i$. We have seen that in this case the use of *a posteriori* auxiliary information led to calibration strategies 2, 3 and 4 in estimator classes 2 and 3. If we wish to take account of the auxiliary information at the sampling stage, it is natural to propose a sampling stratified on x_1 for the first phase and a sampling stratified on x_1 and x_2 for the second phase.

As before, there is no exact parallel between the *a priori* and *a posteriori* procedures, since the use of the information *a priori* mobilizes all the cross-tabulations between the variables x_1 and x_2 .

Thus it is possible to make a choice between incorporating the information either *a priori* or *a posteriori*, and indeed to optimize the sampling plan, when the auxiliary variables are qualitative. The terms of the choice are the same as in a single-phase sampling with a single level of information. An additional consideration is the multiplicity of strata created by the cross-tabulations of x_1 and x_2 in the case in which the modelling used is $y_i = x_{i1}b_1 + x_{i2}b_2 + u_i$, which reinforces the advantages of using the information *a posteriori*.

When the auxiliary variables are quantitative, the choice depends on their conversion into qualitative variables, it not being possible to generalize correctly except by using the parallel between calibration and balanced sampling (cf. Deville 1992).

8. CONCLUSION

In a two-phase sampling, when two different sets of information are available for the total population on the one hand and the sample resulting from the first phase on the other hand, several strategies are possible when one wishes to use the auxiliary information to improve the estimation of totals.

Two different natural approaches have been used to derive estimators: a regression model assisted approach, which seeks to adapt the idea of the regression estimator; and a calibration approach, which attempts to adapt the idea of calibration. The estimators obtained by the two approaches may be linked together. We generated three alternative underlying modellings to which the various estimators obtained may be attached. Thus we obtained

three classes of estimators. Several conceivable calibration strategies were eliminated at the outset as irrelevant.

We have shown that the estimators of a given class, that is, the estimators attached to a given model, are asymptotically equivalent; we gave the form of the variances derived in the case of a linear calibration function, but with asymptotic equivalences, these results remain valid for any calibration function.

For purposes of evaluating strategies, the form of the variances indicates, as intuition would suggest, that one of the classes of estimators (estimators 3, 6 and 7 (calibration strategies 3 and 4)) is preferable to the other from the standpoint of variance when the evaluation is based on the sampling plan alone. When it is based on a modelling of the variable of interest, it suggests that the preferable class of estimators is the one associated with the modelling adopted.

In a situation in which the goal is to adjust a survey and to simultaneously correct the biases that would arise from the use of gross weightings and reduce the variance, the conclusions must be adapted. The changes introduced in the weighting to correct the biases are greater than the corrections to reduce variance. Hence the variables will be incorporated into the calibration once it appears that they affect the probability of selection and thus participate in the creation of bias.

When the auxiliary variables are qualitative, the choice between *a priori* and *a posteriori* use of auxiliary information, that is, between using it at the sampling stage or at the adjustment stage, still rests on the distinction between the two modellings of the variable of interest.

These results may easily be generalized to the case of sampling involving more than two phases.

ACKNOWLEDGEMENTS

I am deeply grateful to Jean-Claude Deville, Louis Meuric and Carl-Erik Särndal for their many helpful suggestions regarding this article.

REFERENCES

- CASSEL, C.M., SÄRNDAL, C.-E., and WRETMAN, J.H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite population. *Biometrika*, 63, 615-620.
- DEVILLE, J.-C. (1992). Constrained samples, conditional inference, weighting: three aspects of the utilisation of auxiliary information. *Proceedings of the Workshop on Uses of Auxiliary Information in Surveys*, October 1992, Örebro.
- DEVILLE, J.-C., and SÄRNDAL, C.-E. (1992). Calibration estimators and generalized raking techniques in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

- DEVILLE, J.-C., SÄRNDAL, C.-E., and SAUTORY, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.
- DEVILLE, J.-C., and DUPONT, F. (1993). Calage et redressement de la non-réponse totale. Journées de Méthodologie.
- DUPONT, F. (1994). Redressements alternatifs en présence de plusieurs niveaux d'information auxiliaire. Working paper of Direction des Statistiques Démographiques et Sociales, F9409.
- FULLER, W.A. (1975). Regression analysis for sample survey. *Sankhyā C*, 37, 117-132.
- GOURIEROUX, C. (1981). *Théorie des sondages*. Edition Economica Paris.
- ISAKI, C.T., and FULLER, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.
- SÄRNDAL, C.-E. (1980). On π inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika*, 67, 639-650.
- SÄRNDAL, C.-E., and SWENSSON, B. (1987). A general view of estimation for two phases of selection with applications to two-phases sampling and nonresponse. *International Statistical Review*, 55, 279-294.
- SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J. (1991). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- WRIGHT, R.L. (1983). Finite population sampling with multivariate auxiliary information. *Journal of the American Statistical Association*, 78, 879-884.

Estimating Some Measures of Income Inequality from Survey Data: An Application of the Estimating Equations Approach

DAVID A. BINDER and MILORAD S. KOVACEVIC¹

ABSTRACT

We summarize some salient aspects of the theory of estimation functions for finite populations. In particular, we discuss the problem of estimation of means and totals and extend this theory to estimating functions. We then apply this estimating functions framework to the problem of estimating measures of income inequality. The resulting statistics are nonlinear functions of the observations. Some of them depend on the order of observations or quantiles. Consequently, the mean squared errors of these estimates are inexpressible by simple formulae and cannot be estimated by conventional variance estimation methods. We show that within the estimating function framework this problem can be resolved using the Taylor linearization method. Finally, we illustrate the proposed methodology using income data from Canadian Survey of Consumer Finance and comparing it to the 'delete-one-cluster' jackknifing method.

KEY WORDS: Complex survey design; Gini family coefficient; Lorenz curve ordinate; Low income measure; Quantile share.

1. INTRODUCTION

The measurement and analysis of economic inequality are well covered in econometrics literature from both, theoretical and applied aspects, although the theoretical issues prevail. Estimation of inequality measures and the impact of the design of sample surveys have gotten less attention. Variance estimation, unavoidable in statistical inference based on these measures, is seldom an issue in the relevant econometric literature. It is usually addressed under very strong assumptions and under unsustainable simplifications of the design or the formulae for the approximate variance. In this paper we present a method that can handle with ease both the estimation of the measures of income inequality and the variance estimation of the resulting non-linear statistics. This method is applicable under a variety of sampling designs.

In general, a population distribution can be described by its cumulative distribution function, $F(y) = \Pr\{Y \leq y\}$, where Y is the random variable corresponding to selecting one population unit at random. Throughout this paper, we assume that Y is non-negative. If Y represents income then we are interested in the properties of an income distribution, such as income concentration, income shares for different population shares, low income proportions, etc. We are also interested in the quantile function $\xi(p) = F^{-1}(p) = \inf\{y \mid F(y) \geq p\}$.

The Lorenz curve, for example, depicts the cumulative income against the population share. The formal definition of the ordinate of the Lorenz curve corresponding to the 100 p -th percentile of the population is

$$L(p) = \frac{\int_0^{\xi_p} y dF(y)}{\mu_Y}, \quad (1.1)$$

where

$$\int_0^{\xi_p} dF(y) = p, \quad \text{and} \quad \int_0^{\infty} y dF(y) = \mu_Y.$$

The finite population form of the expression (1.1), more familiar to survey statisticians, is given by

$$L(p) = \frac{\sum_U Y_i I\{Y_i \leq \xi_p\}}{\sum_U Y_i},$$

where U represents a finite population and $I\{\cdot\}$ is an indicator function.

The income (quantile) share is defined as the percentage of total income shared by the population allocated to the certain income quantile interval $[\xi_{p_1}, \xi_{p_2}]$, $p_1 \leq p_2$. It is equal to the difference of Lorenz curve ordinates

$$Q(p_1, p_2) = L(p_2) - L(p_1).$$

In Figure 1 we give a graph of the Lorenz curve for the Weibull distribution with shape parameter $\alpha = 1.6$, along with the 45° axis. For example, one can read from the graph that not more than 25% of the total income is allocated to the poor half of population, or that the richest 10% of the population earn 20% of the total available income.

¹ David A. Binder, Director, Business Survey Methods Division, and Milorad S. Kovacevic, Senior Methodologist, Household Survey Methods Division, Statistics Canada, R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6.

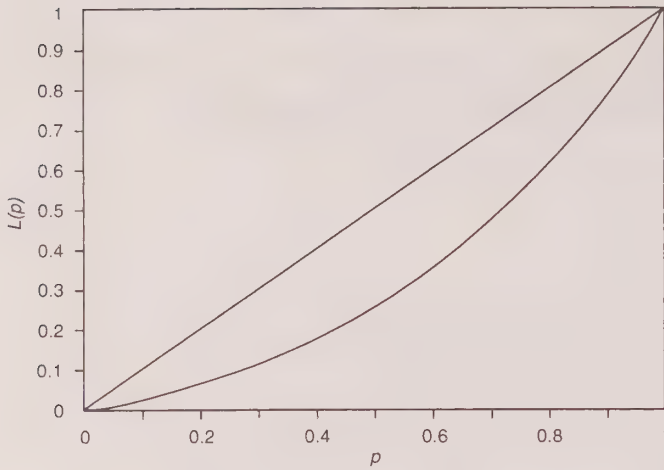


Figure 1. Lorenz Curve for the Weibull Distribution with Shape Parameter $\alpha = 1.6$.

The Gini coefficient measures the degree of the inequality in income distribution. One definition of the Gini coefficient is a linear function of the area between the Lorenz curve and the 45° axis, normalized to lie between 0 and 1. The Gini coefficient in Figure 1 is 0.35. The formal definition of the Gini coefficient (Nygård and Sandström 1981) is

$$G = 1 - 2 \int_0^1 L(p) dp = \frac{1}{\mu} \int_0^\infty [2F(y) - 1] y dF(y).$$

A more general family of Gini coefficients, given in Nygård and Sandström (1981) is

$$G_J = \frac{1}{\mu_Y} \int_0^\infty J[F(y)] y dF(y), \quad (1.2)$$

where J is a bounded and continuous function. For the usual Gini coefficient, $J(p) = 2p - 1$.

Another measure of income inequality used by some economists is the Low Income Measure. This is defined as the proportion of the population units whose income is less than half the median income for the population. Formally, this is

$$\Theta = \int_0^{M/2} dF(y), \quad (1.3a)$$

where M is the median defined by

$$\int_0^M dF(y) = \frac{1}{2}. \quad (1.3b)$$

For all these measures, we can express the parameter of interest, Θ , as the solution to the equation

$$\int u(y, \Theta) dF(y) = 0,$$

where $u(y, \Theta)$ is the kernel of the estimating equation. This estimating equation formulation will be discussed in Section 2. In Sections 3, 4, and 5 we give the estimating equations for the above measures along with the approximation of their mean squared error estimates. In Section 6 we present estimators of these measures based on the complex sample design. Section 7 contains an illustration based on the Canadian Survey of Consumer Finance data.

2. USE OF ESTIMATING EQUATIONS FOR FINITE POPULATIONS

The theory for estimating means and totals from finite populations is now well established in the statistical literature. A formulation which encompasses most estimators used in practice is given in Särndal, Swensson, and Wretman (1992). In this section, we briefly review this theory and show how it can be applied to more complex statistics through the use of estimating equations, as described by Binder (1991) and Binder and Patak (1994).

We begin the exposition of the main idea by reviewing the estimation of the population total T_Y and the finite population distribution function $F(y)$. The estimation of the population total is the core of the estimation equations approach of Binder (1991) and Binder and Patak (1994). Let the population total of the variable Y , be defined as

$$T_Y = N \int y dF(y).$$

Note here that $F(y)$ is a step function corresponding to the distribution function for the finite population. We consider estimators of the form:

$$\hat{T}_Y = \sum_{i \in S} w_i(s) y_i = \sum_{i=1}^N w_i(s) Y_i, \quad (2.1)$$

where $w_i(s)$ is zero whenever the i -th unit is not in the sample. Expression (2.1) gives, for example, the Horvitz-Thompson (HT) unbiased estimator if

$$w_i(s) = \begin{cases} 1/\pi_i, & i \in s, \\ 0, & i \notin s, \end{cases}$$

or the generalized regression estimator if

$$w_i(s) = \begin{cases} [1 + (T_X - \hat{T}_X) x_i / \hat{T}_{X^2}] / \pi_i, & i \in s, \\ 0, & i \notin s, \end{cases}$$

where T_X is the population total of X , and \hat{T}_X and \hat{T}_{X^2} are the HT estimates of the totals of X and X^2 variables, respectively.

Similarly, an estimator for the distribution function is given by

$$N\hat{F}(y) = \sum_{i \in s} w_i(s) I\{y_i \leq y\},$$

where

$$I\{y_i \leq y\} = \begin{cases} 1 & \text{if } y_i \leq y, \\ 0 & \text{if } y_i > y. \end{cases}$$

We note that $\hat{F}(y)$ is uniformly and asymptotically design consistent for $F(y)$, but it is not necessarily a true distribution function, unless

$$\sum_{i \in s} w_i(s) = N.$$

In general, and under certain regularity conditions for complex designs (Francisco and Fuller 1991),

$$\hat{F}(y) - F(y) \rightarrow_p 0, \text{ for any } y.$$

That is, the finite population distribution function, $F(y)$, allows a consistent estimator, $\hat{F}(y)$. This property of the $\hat{F}(y)$ will be used later in proving the consistency of the linearized variance estimators for different income statistics.

Now, we review the application of the estimating equations theory to the estimation of any finite population parameter Θ_o that can be expressed as the solution to

$$\int u(y, \Theta_o) dF(y) = 0.$$

We define the estimating equation estimate for Θ_o as that value of $\hat{\Theta}$ for which

$$\int \hat{u}(y, \hat{\Theta}) d\hat{F}(y) = 0, \quad (2.2)$$

where $\hat{u}(y, \hat{\Theta})$ is an estimate of $u(y, \Theta)$.

We can rewrite (2.2) as

$$\begin{aligned} 0 &= \int \hat{u}(y, \hat{\Theta}) d\hat{F}(y) \\ &= \int [\hat{u}(y, \hat{\Theta}) - u(y, \Theta_o)] d\hat{F}(y) + \int u(y, \Theta_o) d\hat{F}(y) + R, \end{aligned} \quad (2.3)$$

where

$$R = \int [\hat{u}(y, \hat{\Theta}) - u(y, \Theta_o)] [d\hat{F}(y) - dF(y)].$$

The decomposition in (2.3) is the basic starting point for all the derivations of variance in the paper. For each parameter considered we will prove that the remainder term, R , is asymptotically negligible.

Binder (1983) considered the case where $\hat{u}(y, \hat{\Theta}) = u(y, \hat{\Theta})$ and where, for large samples,

$$\begin{aligned} &\int [u(y, \hat{\Theta}) - u(y, \Theta_o)] dF(y) \\ &= (\hat{\Theta} - \Theta_o) \left. \frac{\partial E\{u(y, \Theta)\}}{\partial \Theta} \right|_{\Theta=\Theta_o} + o_p(|\hat{\Theta} - \Theta_o|). \end{aligned}$$

Note that the remainder term R from the decomposition (2.3) should be of order $o_p(|\hat{\Theta} - \Theta_o|)$ to be considered as asymptotically negligible.

For most applications $u(y, \Theta)$ does not need to be estimated by $\hat{u}(y, \hat{\Theta})$. However, for some applications such as the Gini coefficient, the function $u(y, \Theta)$ is estimated so that formula (2.2) allows for these cases in general.

Using these approximations, we have

$$\begin{aligned} \hat{\Theta} - \Theta_o &\approx - \left[\left. \frac{\partial E\{u(y, \Theta)\}}{\partial \Theta} \right|_{\Theta=\Theta_o} \right]^{-1} \\ &\times \int u(y, \Theta_o) d\hat{F}(y) = \int u^*(y) d\hat{F}(y), \quad (2.4) \end{aligned}$$

where

$$u^*(y) = - \left[\left. \frac{\partial E\{u(y, \Theta)\}}{\partial \Theta} \right|_{\Theta=\Theta_o} \right]^{-1} u(y, \Theta_o).$$

Once we have obtained the expression for $u^*(y)$, the derivation of the variance of $\hat{\Theta}$ becomes straightforward. Since we have approximated $\hat{\Theta} - \Theta_o$ as an estimator of a population total of $u^*(y_i)$'s, we can use the mean squared error calculations for the estimate of total to obtain the variance estimate of $\hat{\Theta}$.

For example, for Θ_o equal to the ratio, T_Y/T_X , we have

$$u = y - \Theta_o x,$$

$$u^* = \frac{1}{\mu_X} (y - \Theta_o x).$$

The remainder term in this case is

$$R = \int [y - \hat{\Theta}x - (y - \Theta_0 x)] [d\hat{F}(y) - dF(y)].$$

Therefore,

$$-\frac{R}{\hat{\Theta} - \Theta_0} = [\hat{F}(y) - F(y)]x \rightarrow_p 0,$$

for any y and any finite x .

Similarly, for population quantiles, we have

$$u = I\{y \leq \Theta_0\} - p, \quad (2.5)$$

$$u^* = -\frac{1}{f(\Theta_0)} [I\{y \leq \Theta_0\} - p],$$

where $f(\Theta_0)$ is the value of the density function at Θ_0 . The second expression in (2.5) is an extension of the Bahadur representation for sample quantiles, as described by Francisco and Fuller (1991). Result (2.5) will be used for the ordinates of the Lorenz curve and for the Low Income Measure, which are discussed in Sections 4 and 5.

The remainder term R in this case reduces to $R = \hat{F}(\hat{\Theta}) - \hat{F}(\Theta_0) - F(\hat{\Theta}) + F(\Theta_0)$. In the case of the simple random sample design, Randles (1982) showed that $R = o_p(n^{-1/2})$. For the complex design situation, under some regularity conditions, Shao and Rao (1994) established a similar asymptotic result: first they showed that $\hat{\Theta} - \Theta_0 = O_p(n^{-1/2})$, then that $R = o_p(n^{-1/2})$, and therefore $R = o_p(|\hat{\Theta} - \Theta_0|)$.

3. GINI FAMILY COEFFICIENT

For the Gini family coefficient, given by (1.2), we can use

$$u(y, G_J) = J[F(y)]y - G_J y.$$

Binder's (1983) approach cannot handle the variance estimation of the Gini coefficient. For the Gini coefficient, rather than deriving the variances by breaking the problem into two parts – one for the ratio estimator and the other for the variance of the numerator – we use the estimating equations approach to solve the problem in one step.

Ignoring the remainder term in (2.3), we have the following approximation:

$$0 = \int \{J[\hat{F}(y)]y - \hat{G}_J y\} d\hat{F}(y)$$

$$\approx \int \{J[\hat{F}(y)] - J[F(y)]\} y dF(y) - (\hat{G}_J - G_J) \int y dF(y) + \int \{J[F(y)]y - G_J y\} d\hat{F}(y).$$

Letting

$$\int \{J[\hat{F}(y)] - J[F(y)]\} y dF(y) \approx \int [\hat{F}(y) - F(y)] J'[F(y)] y dF(y),$$

and

$$\begin{aligned} \int \hat{F}(y) J'[F(y)] y dF(y) &= \int \int_0^y J'[F(y)] y d\hat{F}(x) dF(y) \\ &= \int \left[\int_y^\infty J'[F(x)] x dF(x) \right] d\hat{F}(y), \end{aligned}$$

we have that

$$\hat{G}_J - G_J \approx \int u^*(y) d\hat{F}(y),$$

where

$$u^* = \frac{1}{\mu_Y} \left[\int_{F(y)}^1 J'(p) F^{-1}(p) dp + J[F(y)]y - G_J y - E\{F(y)J'[F(y)]y\} \right]. \quad (3.1)$$

For the case of independent and identically distributed observations, this yields the same variance result as described by Glasser (1962) and Sendler (1979). To estimate the variance, it is necessary to use estimates for μ_Y , $F(y)$, and G_J in the expression for u^* .

We investigate the asymptotic behaviour of the remainder term R for the usual Gini coefficient G . The remainder is

$$R = \int \{2y[\hat{F}(y) - F(y)] - y(\hat{G} - G)\} \times [d\hat{F}(y) - dF(y)].$$

Denoting the difference $\hat{F}(y) - F(y)$ by $\bar{D}(y)$, the remainder can be expressed as a sum of two integrals

$$R = \int 2y\hat{D}(y)d\hat{D}(y) - \int (\hat{G} - G)y d\hat{D}(y).$$

The first integral is reduced to zero by the integration by parts, so that the remainder is approximated by

$$\begin{aligned} R &\approx -(\hat{G} - G)(\hat{\mu}_Y - \mu_Y) \\ &= -(\hat{G} - G)o_p(n^{-1/2+\delta}), \quad 0 < \delta < 1/2. \end{aligned}$$

Therefore, we can say that $R = o_p(|\hat{G} - G|)$.

4. LORENZ CURVE ORDINATE AND QUANTILE SHARE

The ordinate of the Lorenz curve was defined in (1.1). In terms of estimating equations, the following two equations are required:

$$\begin{aligned} u_1(y, L(p)) &= I\{y \leq \xi_p\}y - L(p)y, \\ u_2(y) &= I\{y \leq \xi_p\} - p. \end{aligned}$$

The second equation defines the 100 p -th percentile of the distribution; whereas the first equation defines the ordinate of the Lorenz curve in terms of the 100 p -th percentile. Ignoring the remainder term in (2.3), we have the following approximation:

$$\begin{aligned} 0 &= \int [I\{y \leq \hat{\xi}_p\} - \hat{L}(p)]y d\hat{F}(y) \\ &\approx \int_{\xi_p}^{\hat{\xi}_p} y dF(y) - [\hat{L}(p) - L(p)] \int y dF(y) \\ &\quad + \int [I\{y \leq \xi_p\} - L(p)]y d\hat{F}(y). \end{aligned}$$

The first term of this expression can be further approximated as

$$\int_{\xi_p}^{\hat{\xi}_p} y dF(y) \approx (\hat{\xi}_p - \xi_p)\xi_p f(\xi_p),$$

and from (2.5) we see that

$$\hat{\xi}_p - \xi_p \approx - \int \frac{1}{f(\xi_p)} [I\{y \leq \xi_p\} - p] d\hat{F}(y), \quad (4.1)$$

so that

$$(\hat{\xi}_p - \xi_p)\xi_p f(\xi_p) \approx - \int \xi_p [I\{y \leq \xi_p\} - p] d\hat{F}(y).$$

Therefore, to estimate the variance of the ordinate of the Lorenz curve, the appropriate linearization is given by using

$$u^*(y) = \frac{1}{\mu_Y} [(y - \xi_p)I\{y \leq \xi_p\} + p\xi_p - yL(p)].$$

This yields the same result as described by Beach and Davidson (1983) for variances and covariances of ordinates of the Lorenz curve in the case of independent and identically distributed random variables. To estimate the variance it is necessary to use $\hat{\xi}_p$ and $\hat{L}(p)$ in the expression for $u^*(y)$.

To estimate the quantile share $Q(p_1, p_2)$ we need three equations

$$\begin{aligned} u_1(y, Q(p_1, p_2)) &= I\{\xi_{p_1} < y \leq \xi_{p_2}\}y - Q(p_1, p_2)y, \\ u_2(y) &= I\{y \leq \xi_{p_1}\} - p_1, \\ u_3(y) &= I\{y \leq \xi_{p_2}\} - p_2. \end{aligned}$$

Using the same arguments as before, we arrive at

$$\begin{aligned} u^*(y) &= \frac{1}{\mu_Y} [(y - \xi_{p_2})I\{y \leq \xi_{p_2}\} \\ &\quad - (y - \xi_{p_1})I\{y \leq \xi_{p_1}\} \\ &\quad + p_2\xi_{p_2} - p_1\xi_{p_1} - yQ(p_1, p_2)]. \end{aligned}$$

5. LOW INCOME MEASURE

The Low Income Measure was defined in (1.3). In terms of estimating equations, the following two equations are required:

$$\begin{aligned} u_1(y, \Theta) &= I\left\{y \leq \frac{M}{2}\right\} - \Theta, \\ u_2(y) &= I\{y \leq M\} - \frac{1}{2}, \end{aligned}$$

where M denotes the median of the distribution defined by the second equation, whereas the first equation defines the Low Income Measure in terms of the median. Ignoring the remainder term in (2.3), we have the following approximation:

$$\begin{aligned}
0 &= \int \left(I\left\{y \leq \frac{\hat{M}}{2}\right\} - \hat{\Theta} \right) d\hat{F}(y) \\
&\approx \frac{1}{2} (\hat{M} - M) f\left(\frac{M}{2}\right) - (\hat{\Theta} - \Theta) \\
&\quad + \int \left(I\left\{y \leq \frac{M}{2}\right\} - \Theta \right) d\hat{F}(y).
\end{aligned}$$

Using result (4.1) to substitute for $\hat{M} - M$, and solving for $\hat{\Theta} - \Theta$, we obtain

$$\hat{\Theta} - \Theta \approx \int u^*(y) d\hat{F}(y),$$

where

$$\begin{aligned}
u^* &= - \frac{f\left(\frac{M}{2}\right)}{2f(M)} \left(I\{y \leq M\} - \frac{1}{2} \right) \\
&\quad + I\left\{y \leq \frac{M}{2}\right\} - \Theta. \quad (5.1)
\end{aligned}$$

The problem with applying this result to estimate the variance of the estimated Low Income Measure is that it is necessary to estimate $f(M)$ and $f(M/2)$. To accomplish this, we could use

$$\hat{f}(\xi) = \frac{\hat{F}\left(\xi + \frac{h}{2}\right) - \hat{F}\left(\xi - \frac{h}{2}\right)}{h},$$

for some suitably small h . Alternatively, we could perform the following calculations, as suggested by Francisco and Fuller (1991) for another problem. For a given value of ξ , we estimate the corresponding percentile, $100p$. We then construct the Woodruff interval for that percentile. This is determined by first solving for h_1 and h_2 in

$$\begin{aligned}
\inf_{h_1} \left[\frac{\int [I\{y \leq \xi - h_1\} - p] d\hat{F}(y)}{\left[\text{mse} \left\{ \int [I\{y \leq \xi\} - p] d\hat{F}(y) \right\} \right]^{1/2}} \right] &\leq -z_{1-\alpha/2}, \\
\inf_{h_2} \left[\frac{\int [I\{y \leq \xi + h_2\} - p] d\hat{F}(y)}{\left[\text{mse} \left\{ \int [I\{y \leq \xi\} - p] d\hat{F}(y) \right\} \right]^{1/2}} \right] &\geq z_{1-\alpha/2},
\end{aligned}$$

where $z_{1-\alpha/2}$ is the $100(1 - \alpha/2)$ -th percentile from the standard normal distribution. Then we compute

$$\hat{f}(\xi) = \frac{2z_{1-\alpha/2} \left[\text{mse} \left\{ \int [I\{y \leq \xi\} - p] d\hat{F}(y) \right\} \right]^{1/2}}{h_1 + h_2}. \quad (5.2)$$

This calculation uses the asymptotic equivalence of $\hat{\xi} - \xi$ and the estimated sum of the $u^*(y)$'s given by (2.5).

We see that the estimated variance for the Low Income Measure may be somewhat complex to compute. The estimating functions framework has however provided us with the appropriate formulae.

The discussion about the remainder term in the decomposition (2.3) of the low income measure is analogous to that made for the case of the quantile estimation (2.5).

6. ESTIMATION WITH A COMPLEX SURVEY

Let us assume a stratified multistage design with a large number of strata, H , with a few primary sampling units (clusters), n_h (≥ 2), sampled from each stratum. For example, in the Canadian Survey of Consumer Finance (SCF) which uses the Labour Force Survey (LFS) vehicle, the number of strata is several hundreds and the number of clusters per stratum is on average less than six. Let w_{hci} be the normalized weight attached to the i -th ultimate unit in the c -th cluster of the h -th stratum such that the appropriate estimator of mean and the consistent estimator of its mean squared error are

$$\hat{\mu} = \sum_s w_{hci} y_{hci}$$

$$\text{mse}(\hat{\mu}) = \sum_h \frac{n_h}{n_h - 1} \sum_c (u_{hc}^* - \bar{u}_h^*)^2 \quad (6.1)$$

where $u_{hc}^* = \sum_i w_{hci}(y_{hci} - \hat{\mu})$ and $\bar{u}_h^* = 1/n_h \sum_c u_{hc}^*$. We use $\sum_s = \sum_h \sum_c \sum_i$ to denote summation over all ultimate units in the sample incorporating all stages of sampling. We assume that PSU's are selected with replacement.

This paper is not concerned with the efficiency of the estimators but rather the properties of commonly used estimators. An analysis of more complex estimators found in the econometric literature is beyond the scope of our study.

An estimator of the finite population distribution function is

$$\hat{F}(y) = \sum_s w_{hci} I\{y_{hci} \leq y\}.$$

A consistent estimator of the approximation of the mean squared error of the distribution function estimated in y takes the form (6.1) where $u_{hc}^* = \sum_i w_{hci} [I\{y_{hci} \leq y\} - \hat{F}(y)]$.

The usual estimate of the finite population quantile is the sample quantile

$$\hat{\xi}_p = \inf\{y_{hci} \in S : \hat{F}(y_{hci}) \geq p\}$$

which is the solution of the estimating equation

$$\sum_s w_{hci} [I\{y_{hci} \leq \hat{\xi}_p\} - p] = 0.$$

Accordingly, using result (2.5), the estimator of the mean squared error of the p -th quantile has the form (6.1) with

$$u_{hc}^* = \frac{1}{[\hat{f}(\hat{\xi}_p)]^2} \sum_i w_{hci} [I\{y_{hci} \leq \hat{\xi}_p\} - p].$$

If the expression (5.2) is used for the estimation of the density function $f(\xi)$, the MSE estimate of the quantile $\hat{\xi}_p$ becomes

$$\text{mse}_\alpha(\hat{\xi}_p) = \left(\frac{D_\alpha(\hat{\xi}_p)}{z_{1-\alpha/2}} \right)^2 \quad (6.2)$$

where $D_\alpha(\hat{\xi}_p) = (h_1 + h_2)/2 = (\hat{\xi}_U - \hat{\xi}_L)/2$ is the half length of the $100(1 - \alpha)\%$ confidence interval for $\hat{\xi}_p$. In a complex sample design, h_1 and h_2 are obtained as solutions of

$$\hat{\xi}_L = \hat{\xi}_p - h_1 =$$

$$\inf\{y_{hci} \in S : \hat{F}(y_{hci}) \geq p - z_{1-\alpha/2} \sqrt{\text{mse}[\hat{F}(\hat{\xi}_p)]}\}$$

$$\hat{\xi}_U = \hat{\xi}_p + h_2 =$$

$$\inf\{y_{hci} \in S : \hat{F}(y_{hci}) \geq p + z_{1-\alpha/2} \sqrt{\text{mse}[\hat{F}(\hat{\xi}_p)]}\}.$$

The estimator (6.2) was also used by Francisco and Fuller (1991). Generally speaking the motivation for (5.2) and consequently for (6.2) comes from Woodruff's (1952) confidence interval for individual quantiles. Francisco and Fuller (1986) and Rao and Wu (1987) used these intervals to derive variance estimators. Although the estimator depends on the confidence coefficient, they showed that it is asymptotically consistent for any significance level α . Rao and Wu (1987) studied the standard errors of quantiles for the cluster samples estimated in this manner. Their Monte Carlo results suggest that 95% confidence interval works well as a basis for extracting the standard error. Binder and Patak (1994) obtained a similar form of the

variance estimator by using the estimating equations approach.

The estimate of the usual Gini coefficient is the solution of the following estimating equation

$$\sum_s w_{hci} \{ [2\hat{F}(y_{hci}) - 1] y_{hci} - \hat{G} y_{hci} \} = 0$$

and takes the form

$$\hat{G} = \frac{2}{\hat{\mu}} \sum_s w_{hci} \hat{F}(y_{hci}) y_{hci} - 1$$

where $\hat{\mu} = \sum_s w_{hci} y_{hci}$.

The estimate of the MSE of the Gini coefficient can be computed using expression (6.1) by replacing u_{hc}^* , originally defined by (3.1), with its complex survey form. After some algebraic manipulation we obtain the following expression:

$$u_{hc}^* = \frac{2}{\hat{\mu}} \sum_i w_{hci} \left[A(y_{hci}) y_{hci} + B(y_{hci}) - \frac{\hat{\mu}}{2} (\hat{G} + 1) \right]$$

where

$$A(y) = \hat{F}(y) - \frac{\hat{G} + 1}{2}$$

and

$$B(y) = \sum_s w_{hci} y_{hci} I\{y_{hci} \geq y\}.$$

The Lorenz curve ordinates could be obtained by solving a system of estimating equations

$$\sum_s w_{hci} [I\{y_{hci} \leq \hat{\xi}_p\} y_{hci} - \hat{L}(p) y_{hci}] = 0$$

$$\sum_s w_{hci} [I\{y_{hci} \leq \hat{\xi}_p\} - p] = 0.$$

The resulting estimate is

$$\hat{L}(p) = \frac{1}{\hat{\mu}} \sum_s w_{hci} y_{hci} I\{y_{hci} \leq \hat{\xi}_p\}.$$

To estimate the mean squared error of the Lorenz curve ordinates we simply use the values of u_{hc}^* defined by (6.3) in (6.1)

$$u_{hc}^* = \frac{1}{\hat{\mu}} \sum_i w_{hci} [(y_{hci} - \hat{\xi}_p) I\{y_{hci} \leq \hat{\xi}_p\} + p \hat{\xi}_p - y_{hci} \hat{L}(p)]. \quad (6.3)$$

Similarly, the mse of the quantile share

$$\hat{Q}(p_1, p_2) = \frac{1}{\hat{\mu}} \sum_s w_{hci} y_{hci} I\{\hat{\xi}_{p_1} < y_{hci} \leq \hat{\xi}_{p_2}\}$$

is approximated by (6.1) using

$$\begin{aligned} u_{hc}^* = \frac{1}{\hat{\mu}} \sum_i w_{hci} [& (y_{hci} - \hat{\xi}_{p_2}) I\{y_{hci} \leq \hat{\xi}_{p_2}\} \\ & - (y_{hci} - \hat{\xi}_{p_1}) I\{y_{hci} \leq \hat{\xi}_{p_1}\} \\ & + p_2 \hat{\xi}_{p_2} - p_1 \hat{\xi}_{p_1} - y_{hci} \hat{Q}(p_1, p_2)]. \end{aligned}$$

The Low Income Measure defined by (1.3) is estimated as

$$\hat{\Theta} = \hat{F}(\hat{M}/2) = \sum_s w_{hci} I\{y_{hci} \leq \hat{M}/2\}.$$

The mean squared error of the low income measure can be estimated approximately by the expression (6.1), where, (from the equation (5.1)):

$$\begin{aligned} u_{hc}^* = - \frac{\hat{f}(\hat{M}/2)}{2\hat{f}(\hat{M})} \sum_i w_{hci} [& I\{y_{hci} \leq \hat{M}\} - 1/2] \\ & + \sum_i w_{hci} [I\{y_{hci} \leq \hat{M}/2\} - \hat{\Theta}]. \end{aligned}$$

7. ILLUSTRATION

The methodology above is illustrated with an application to the family income data collected in the Canadian Survey of Consumer Finance (SCF). We use the file on the Disposable Income of Economic Families obtained for the province of Ontario in 1988. Disposable income is defined as total income after tax reported in the survey. The SCF uses the framework of the Canadian Labour Force Survey which is based on a stratified, multistage design. For more details on the sample design see Singh *et al.* (1990).

We estimated the median \hat{M} , the Gini coefficient \hat{G} , the Low Income Measure $\hat{\Theta}$, Lorenz Curve Ordinates and quintile shares $Q(0, .2)$, $Q(.2, .4)$, $Q(.4, .6)$, $Q(.6, .8)$, $Q(.8, 1.0)$. Their standard errors are obtained using the proposed methodology and the jackknife 'delete-one-cluster' method.

We present a brief description of the jackknife 'delete-one-cluster' method used for this illustration. First, we assume that the estimate of the unknown parameter Θ can be expressed as $\hat{\Theta} = \mathcal{L}(\hat{F})$, where \hat{F} is the estimated distribution function. The estimate of the distribution function $\hat{F}_{(hj)}$ obtained from the sample after removing

the j -th sampled cluster of the h -th stratum ($j = 1, \dots, n_h$, $h = 1, \dots, H$) is

$$\hat{F}_{(gj)}(y) = \sum_s A_{hci}(g, j) w_{hci} I\{y_{hci} \leq y\}$$

$$\text{where } A_{hci}(g, j) = \begin{cases} 1, & h \neq g; \\ \frac{n_g}{n_g - 1}, & h = g, c \neq j; \\ 0, & h = g, c = j. \end{cases}$$

Then $\hat{\Theta}_{(gj)} = \mathcal{L}(\hat{F}_{(gj)})$ and the resulting 'delete-one-cluster' jackknife estimator of the variance of $\hat{\Theta} = \mathcal{L}(\hat{F})$ is

$$\text{var}_J(\hat{\Theta}) = \sum_{g=1}^H \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} (\hat{\Theta}_{(gj)} - \hat{\Theta})^2.$$

It is known that the jackknife variance estimator performs poorly for quantiles due to its inconsistency (Kovar *et al.* 1988). There are some recent results (Shao and Wu 1989, Rao, Wu and Yue 1992) that suggest that the 'delete d ' jackknife and 'delete-one-cluster', under certain conditions, may have desirable asymptotic properties for the variance estimation of non-smooth statistics like quantiles or the low income measure. On the other hand, for statistics like the Gini coefficient the jackknife estimator of the asymptotic variance is consistent (Shao 1993).

Unlike jackknifing, the estimating equations approach is not computationally intensive. It is simple, explicit and incorporates the sample design. It provides formulae for the asymptotic variance that are easy to program despite their complicated form.

Realizing the limitations imposed by using a single sample to make an objective comparison between different methods, the purpose of this example is to point out differences in the standard errors obtained by the estimating equations approach and a computationally intensive method like the jackknifing. Results are summarized in the table below. The direction of the difference in the estimated standard errors confirms the overall conservativeness of the jackknifing method. The difference can be attributed to the upward bias of the jackknifing method in the case of the median, although the 'delete-one-cluster' jackknife is preferable to the 'delete-1' jackknife. For the quantile shares it can be partly explained by the fact that upper quantile shares may not cut over all primary sampling units but rather perform as separated classes which may affect the jackknifing more than the estimating equations method.

Table 1

Measures of Income Inequality and Their Standard Errors

Measure	Estimate	Standard Error	
		Estimating Equations Approach	Jackknifing 'Delete-One-Cluster'
Median	31705	303.3	569.8
Gini	0.3482	0.005	0.005
Low Income Measure	0.1980	0.00586	0.00613
Lorenz Curve Ordinates			
L(0.2)	0.0561	0.00137	0.00175
L(0.4)	0.1745	0.00166	0.00194
L(0.6)	0.3522	0.00246	0.00285
L(0.8)	0.5982	0.00317	0.00393
Quintile Shares			
Q(0, 0.2)	0.0561	0.00137	0.00167
Q(0.2, 0.4)	0.1186	0.00159	0.00221
Q(0.4, 0.6)	0.1775	0.00157	0.00282
Q(0.6, 0.8)	0.2461	0.00158	0.00337
Q(0.8, 1.0)	0.4017	0.00395	0.00451

8. SUMMARY

The problem of estimating the variance of complex statistics, such as measures of income inequality, have eluded statisticians for years. Replication methods such as the jackknife are often suggested for estimation. The advantage of the linearization approach is that it can be used under a wide class of sampling designs and does not suffer from the need for intensive computations which methods such as the bootstrap entail. Through the method of estimating functions and the decomposition given in (2.3), we find that some difficult problems can be solved more easily. A discussion about the order of the remainder term for some of these measures is given as well. A more rigorous proof for a complex sample design can be established along the lines given in Shao and Rao (1994).

REFERENCES

- BEACH, C.M., and DAVIDSON, R. (1983). Distribution-free statistical inference with Lorenz curves and income shares. *Review of Economic Studies*, 50, 723-735.
- BINDER, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.
- BINDER, D.A. (1991). Use of estimating functions for interval estimation from complex surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 34-42.
- BINDER, D.A., and PATAK, Z. (1994). Use of estimating functions for interval estimation from complex surveys. *Journal of the American Statistical Association*, 89, 1035-1044.
- FRANCISCO, C.A., and FULLER, W.A. (1986). Estimation of the Distribution function with a complex survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 37-45.
- FRANCISCO, C.A., and FULLER, W.A. (1991). Quantile estimation with a complex survey design. *Annals of Statistics*, 19, 454-469.
- GLASSER, G.J. (1962). Variance formulas for the mean difference and coefficient of concentration. *Journal of the American Statistical Association*, 57, 648-654.
- KOVAR, J.G., RAO, J.N.K., and WU, C.F.J. (1988). Bootstrap and other methods to measure errors in survey estimates. *The Canadian Journal of Statistics*, 16, 25-45.
- NYGÅRD, F., and SANDSTRÖM, A. (1981). *Measuring Income Inequality*. Stockholm: Almqvist and Wiksell International.
- RANDLES, R.H. (1982). On the Asymptotic Normality of Statistics with Estimated Parameters. *Annals of Statistics*, 10, 462-474.
- RAO, J.N.K., and WU, C.F.J. (1987). Methods for Standard Errors and Confidence Intervals from Survey Data: Some Recent Work. *Proceedings of the 46th session, International Statistical Institute*, 3, 5-19.
- RAO, J.N.K., WU, C.F.J., and YUE, K. (1992). Some Recent Work on Resampling Methods for Complex Surveys. *Survey Methodology*, 18, 209-217.
- SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- SENDER, W. (1979). On statistical inference in concentration measurement. *Metrika*, 26, 109-122.
- SHAO, J., and RAO, J.N.K. (1994). Standard Errors for Low Income Proportions Estimated from Stratified Multi-Stage Samples. *Sankhyā, B*, (to appear).
- SHAO, J. and WU, C.W.J. (1989). A general Theory for Jackknife Variance Estimation. *Annals of Statistics*, 17, 1176-1197.
- SHAO, J. (1993). Inferences Based on *L*-statistics in Survey Problems: Lorenz Curve, Gini Family and Poverty Proportion. In *Proceedings of the Workshop on Statistical Issues in Public Policy Analysis*, Carleton University and University of Ottawa.
- SINGH, M.P., DREW, J.D., GAMBINO, J.G., and MAYDA, F. (1990). *Methodology of the Canadian Labour Force Survey*, Catalogue No. 71-526, Statistics Canada.
- WOODRUFF, R.S. (1952). Confidence intervals for medians and other position measures. *Journal of the American Statistical Association*, 47, 635-646.

A Reduced-Size Transportation Algorithm for Maximizing the Overlap Between Surveys

LAWRENCE R. ERNST and MICHAEL M. IKEDA¹

ABSTRACT

When redesigning a sample with a stratified multi-stage design, it is sometimes considered desirable to maximize the number of primary sampling units retained in the new sample without altering unconditional selection probabilities. For this problem, an optimal solution which uses transportation theory exists for a very general class of designs. However, this procedure has never been used in the redesign of any survey (that the authors are aware of), in part because even for moderately-sized strata, the resulting transportation problem may be too large to solve in practice. In this paper, a modified reduced-size transportation algorithm is presented for maximizing the overlap, which substantially reduces the size of the problem. This reduced-size overlap procedure was used in the recent redesign of the Survey of Income and Program Participation (SIPP). The performance of the reduced-size algorithm is summarized, both for the actual production SIPP overlap and for earlier, artificial simulations of the SIPP overlap. Although the procedure is not optimal and theoretically can produce only negligible improvements in expected overlap compared to independent selection, in practice it gave substantial improvements in overlap over independent selection for SIPP, and generally provided an overlap that is close to optimal.

KEY WORDS: Linear programming; Sample redesign; Survey of Income and Program Participation.

1. INTRODUCTION

The problem of maximizing the expected number of primary sampling units (PSUs) retained in sample when redesigning a survey with a stratified design for which the PSUs are selected with probability proportional to size was introduced to the literature by Keyfitz (1951). Typically, the motivation for maximizing the overlap of PSUs is to reduce additional costs, such as the training of a new interviewer for a household survey, incurred with each change of sample PSU. Procedures for maximizing overlap do not alter the unconditional probability of selection for a set of PSUs in a new stratum, but conditions its probability of selection in such a manner that the probability of a PSU being selected in the new sample is generally greater than its unconditional probability when the PSU was in the initial sample and less otherwise.

Overlap procedures are applicable when the redesign results in either a restratification of the PSUs or a change in their selection probabilities. Keyfitz (1951) presented an optimal procedure, but only for one-PSU-per-stratum designs in the special case when the initial and new strata are identical, with only the selection probabilities changing. Causey, Cox and Ernst (1985) obtained an optimal solution to the overlap problem under very general conditions by formulating it as a transportation problem, which is a special form of linear programming problem. This procedure imposes no restrictions on changes in strata definitions or number of PSUs per stratum. (A similar result had

been independently obtained by Arthanari and Dodge (1981), although they did not discuss the issue of changes in strata definitions. Both sets of authors obtained their results by generalizing work of Raj (1968).) However, there are at least two other difficulties with the procedure of Causey, Cox and Ernst which can make it unusable in practice, one which is the focus of Ernst (1986), and the other the focus of the current paper.

The first difficulty is that, if the initial sample of PSUs was not selected independently from stratum to stratum, the information necessary to compute all the joint probabilities required by this method may not be available in practice. An alternative linear programming procedure, for use in such cases, was developed by Ernst (1986). The Bureau of the Census has used linear programming to overlap its demographic surveys on five occasions. On four of these occasions (the selection of the 1980s and 1990s Current Population Survey (CPS) designs, and the 1980s and 1990s National Crime Victimization Survey (NCVS) designs) the procedure in Ernst (1986) was used because the initial design was not selected independently from stratum to stratum. In particular, as explained in Ernst (1986), if the initial sample was itself selected by overlapping with a still earlier design then this independence assumption generally does not hold, which was the key reason why it did not hold for these four redesigns.

The second difficulty with the optimal procedure is that the transportation problem may be too large to solve in practice. The Bureau of the Census also used linear

¹ Lawrence R. Ernst, Chief, Research Group, Office of Compensation and Working Conditions, Bureau of Labor Statistics, Washington, DC 20212, U.S.A.; Michael M. Ikeda, Mathematical Statistician, Statistical Research Division, Bureau of the Census, Washington, DC 20233, U.S.A.

programming to overlap the 1990s Survey of Income and Program Participation (SIPP) design with the 1980s SIPP design, both two-PSUs-per-stratum designs. The initial sample for SIPP was selected independently from stratum to stratum. However, the transportation problem for the optimal procedure would have been too large to practically solve for many strata. This is because for each new stratum to be overlapped consisting of n PSUs, the number of variables in the transportation problem for the optimal procedure can be as large as $2^n \times \binom{n}{2}$. The largest value of n for which a transportation problem with that many variables can be solved with the computer facilities that we have used is approximately $n = 15$.

This paper presents a reduced-size formulation of the overlap procedure as a transportation problem which decreases the numbers of variables in the SIPP problem to $((\binom{n}{2} + n + 1) \times \binom{n}{2})$, a striking reduction for moderate to large values of n . The procedure assumes that the initial sample was selected independently from stratum to stratum, and hence could not have been used instead of the procedure of Ernst (1986) to overlap the CPS and NCVS designs. This reduced-size procedure has been successfully run for strata with as many as 68 PSUs. In contrast, for $n = 68$, the $2^{68} \times \binom{68}{2}$ possible number of variables for the unreduced formulation is far beyond the size of problem that can be solved by any current computer. Furthermore, though the reduced-size procedure sacrifices optimality in exchange for its size reduction, it does appear in practice to yield results fairly close to optimal, as we will show. The reduced-size procedure is the procedure that was used to overlap SIPP.

In Section 2 the procedure of Causey, Cox and Ernst (1985) is reviewed, to provide background for the presentation of the reduced-size procedure.

The reduced-size procedure is presented in Section 3. Although the approach has general applicability, for ease of presentation it is only described in detail for the case when both the initial and new designs are two-PSUs-per-stratum without replacement. A small, artificial example of the reduced-size procedure is also presented in Section 3. This example serves to illustrate the procedure and to demonstrate that the ordering of the pairs of PSUs in a new design stratum, a key step in the algorithm, affects the expected overlap. We also outline in this section some analytical results on the comparison between the reduced-size procedure and the optimal procedure. Upper bounds on the loss in expected overlap from using the reduced-size procedure instead of the optimal procedure are stated. It is also explained that in certain situations this loss can approach two PSUs for two-PSUs-per-stratum designs, the worst possible situation. Further details and proofs of the results in this section as well as some results in other sections are presented in Ernst and Ikeda 1994.

In Section 4 the performance of the reduced-size procedure is presented, both for the actual SIPP production

overlap and for earlier, artificial simulations of the SIPP overlap. The expected overlap for this procedure is compared to that for independent selection of the new sample PSUs and to an upper bound on the optimal expected overlap. The results show that for this application, in contrast with some of the theoretical results described in Section 3, the expected overlap with the reduced-size procedure is much larger than if independent selection had been used to select the new sample PSUs, and nearly as large as the optimal expected overlap. Also presented are computer running times for the reduced-size procedure as a function of stratum size.

Finally, our conclusions are stated in Section 5.

2. REVIEW OF THE OVERLAP PROCEDURE OF CAUSEY, COX AND ERNST (1985)

The overlap procedure of Causey, Cox and Ernst (1985), like all overlap procedures, conditions the selection of sample PSUs in each new stratum in some way on which PSUs in the stratum were in the initial sample. This particular overlap procedure attains true optimality by making complete use of this information and formulating the procedure as a transportation problem. We proceed to present this procedure.

First, however, we introduce some notation that will be used throughout the paper. Let S denote a stratum in the new design. Each such stratum corresponds to a separate overlap problem. Let n denote the number of PSUs in S and let A_1, \dots, A_n denote the PSUs in S . Let I denote the random subset of $\{1, \dots, n\}$ such that $k \in I$ if and only if A_k was in the initial sample, and let N denote the corresponding set with respect to the new sample. For example, if A_2 and A_3 were the PSUs in S that were in the initial sample and A_1 and A_3 are the PSUs in the new sample, then $I = \{2, 3\}$ and $N = \{1, 3\}$. Let m^* , n^* denote the number of possible values for I and N , respectively. Let J_i , $i = 1, \dots, m^*$, denote the possible values for I and let S_j , $j = 1, \dots, n^*$, denote the possible values for N . The goal of all overlap procedures is to maximize the expected number of PSUs in $N \cap I$, while preserving the values of the $P(S_j)$'s.

To illustrate some of these concepts further, consider an example for which $n = 3$. Then $n^* = 3$ if the new design is either 1 or 2 PSUs per stratum with the values for N , that is the S_j 's, consisting of $\{1\}, \{2\}, \{3\}$ in the 1 PSU per stratum case and $\{1, 2\}, \{1, 3\}, \{2, 3\}$ in the two PSUs per stratum case. Suppose PSUs A_1 and A_2 were in one initial stratum and PSU A_3 was in another initial stratum and there were three PSUs in each of these initial strata. If the initial design was 1 PSU per stratum, then $m^* = 6$, with the values of I , that is the J_i 's, consisting of $\emptyset, \{1\}, \{2\}, \{3\}, \{1, 3\}, \{2, 3\}$; if the initial design was 2 PSUs per stratum then $m^* = 6$, with the J_i 's consisting of $\{1\}, \{2\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}$.

We now present the transportation problem for the overlap procedure of Causey, Cox and Ernst (1985). Abbreviate by $P(J_i)$ the probability that $I = J_i$ and by $P(S_j)$ the probability that $N = S_j$. In addition, let x_{ij} be the variable denoting the joint probability of these two events, and let c_{ij} denote the number of elements in $J_i \cap S_j$. The $P(J_i)$'s, $P(S_j)$'s and c_{ij} 's are known values, while the x_{ij} 's are variables for which the optimal values are to be determined. Then the transportation problem to solve is to determine $x_{ij} \geq 0$ which maximize

$$\sum_{i=1}^{m^*} \sum_{j=1}^{n^*} c_{ij} x_{ij} \quad (2.1)$$

subject to

$$\sum_{j=1}^{n^*} x_{ij} = P(J_i), \quad i = 1, \dots, m^*, \quad (2.2)$$

$$\sum_{i=1}^{m^*} x_{ij} = P(S_j), \quad j = 1, \dots, n^*. \quad (2.3)$$

Note that in this transportation problem, the objective function (2.1) is the expected number of PSUs in S that are in $N \cap I$. Also note that the constraints (2.2) and (2.3) are required by the definitions of the $P(J_i)$'s, $P(S_j)$'s and the x_{ij} 's.

Once the optimal x_{ij} 's have been obtained, the conditional probability that $N = S_j$ given that $I = J_i$ is then $x_{ij}/P(J_i)$ for all i, j .

We present an example to illustrate the use of the formulation (2.1)–(2.3) in the case where both the initial and new designs are two-PSUs-per-stratum without replacement. In this example, and throughout the paper, p_i, π_i denote the predetermined probability that $i \in I$ and $i \in N$, respectively.

Consider a final stratum S with $n = 3$. All of the PSUs were in different initial strata. Let $p_1 = .6, p_2 = .75, p_3 = .7, \pi_1 = .5, \pi_2 = .8, \pi_3 = .7$. Since the PSUs were all in different initial strata, there are 8 different possibilities for I , with probabilities given in Table 1.

Table 1

Probabilities for Possible Sets of Initial Sample PSUs

i	1	2	3	4	5	6	7	8
J_i	{1,2,3}	{1,2}	{1,3}	{2,3}	{1}	{2}	{3}	\emptyset
$P(J_i)$.315	.135	.105	.21	.045	.09	.07	.03

Since the new design is two-PSUs-per-stratum without replacement, there are 3 different possibilities for N ,

namely the pairs $S_1 = \{1,2\}, S_2 = \{1,3\}, S_3 = \{2,3\}$, and hence $P(S_1) = .30, P(S_2) = .20, P(S_3) = .50$.

Furthermore, the values of c_{ij} are then as given in Table 2. Upon maximizing (2.1) subject to (2.2) and (2.3) with the given $P(J_i)$'s, $P(S_j)$'s and c_{ij} 's, an optimal set of x_{ij} 's, presented in Table 2, is obtained. Finally, by dividing each of the x_{ij} entries in row i of Table 2 by $P(J_i)$, an optimal set of conditional probabilities $P(S_j | J_i)$, is obtained. For example, since $x_{12} = .025$ and $P(J_1) = .315$, it follows that $P(S_2 | J_1) = 5/63$.

Table 2

Values of c_{ij} and Values of x_{ij} that Maximize Overlap for Optimal Procedure

i	c_{ij}			x_{ij}		
	j			j		
	1	2	3	1	2	3
1	2	2	2	.000	.025	.290
2	2	1	1	.135	.000	.000
3	1	2	1	.000	.105	.000
4	1	1	2	.000	.000	.210
5	1	1	0	.045	.000	.000
6	1	0	1	.090	.000	.000
7	0	1	1	.000	.070	.000
8	0	0	0	.030	.000	.000

For this example, as can be computed from (2.1) and Table 2, the expected overlap under the optimal procedure is 1.735 PSUs. In comparison, the expected overlap if the initial and final designs are selected independently is $p_1\pi_1 + p_2\pi_2 + p_3\pi_3 = 1.39$ PSUs.

For two-PSU-per-stratum without replacement problems, the possible values for N are always the $\binom{n}{2}$ subsets of $\{1, \dots, n\}$ of size 2, that is $n^* = \binom{n}{2}$. However m^* can vary widely. $m^* = \binom{n}{2}$ when the PSUs in S comprise a single initial stratum. The upper bound of 2^n on m^* is attained when all the PSUs in S were in different initial strata, as illustrated by the previous example, and in some other situations. A general, exact expression for m^* is presented in Ernst and Ikeda (1994).

For the two-PSUs-per-stratum without replacement overlap problem, the number of variables in the transportation problem for the optimal procedure is m^*n^* which can be as large as $2^n \binom{n}{2}$. For $n = 15$, $2^n \binom{n}{2} = 3,440,640$, which is about as large a transportation problem as can be solved with the computer facilities that we used. However, $n > 15$ for nearly half the nonselfrepresenting strata (that is strata consisting of noncertainty PSUs) in our SIPP application, and consequently it was necessary to develop a procedure, described in the next section, which reduces the size of the transportation problem, while still producing nearly maximal expected overlap in practice.

3. THE ALGORITHM FOR THE REDUCED-SIZE PROCEDURE

Previous work on reducing the size of the transportation problem (2.1)–(2.3) has focused on accomplishing the size reduction while retaining optimality. For example, the approach of Aragon and Pathak (1990) retains optimality and reduces the size of the problem by 75 percent when $m^* = n^*$. Unfortunately, when m^* is much larger than n^* , which is when size reduction is most needed, their method produces negligible size reduction in relative terms. A generalization of this approach is presented in Pathak and Fahimi (1992), but there is no indication that their procedure always yields a size reduction that is substantial in relative terms.

In this section a reduced-size procedure is presented which takes a different approach. We sacrifice optimality, at least in theory, in return for an assured size reduction down to a manageable size transportation problem. This size reduction is accomplished, in the case when the initial and new designs are both two PSUs per stratum for example, by ordering all pairs of PSUs in a new stratum and then conditioning the new selection probabilities for any initial set of sample PSUs of size greater than 2 on the first pair of PSUs in the ordering contained in the initial set, rather than conditioning on the entire initial set. That is, each possible initial set of sample PSUs which consists of more than 2 PSUs is combined with a set of size 2. As illustrated in Section 4, this procedure may yield a near optimal overlap in practice; particularly with an appropriate ordering of the pairs of PSUs, as described in Section 3.1.2.

The reduced-size procedure is applicable whenever PSUs in the initial and new designs are selected without replacement. However, the procedure will be described in detail, in Section 3.1, only for the case when both the initial and new designs are two-PSUs-per-stratum. Then, in Section 3.2, the changes necessary to apply this procedure for other initial and new designs will be sketched. Finally, in Section 3.3, some analytical results are outlined on the relationships among the expected overlap for the reduced-size procedure, the optimal procedure and independent selection. It is assumed throughout this section that PSUs in the initial sample were selected independently from stratum to stratum.

3.1 Reduced-Size Procedure When Both Designs Are Two-PSUs-Per-Stratum

The reduced-size procedure to be described includes the following key aspects: the specific ordering of the pairs of PSUs; the reformulation of the transportation problem (2.1)–(2.3) for the reduced size procedure; the computation of the probabilities for the initial outcomes for this formulation; and the computation of the cost coefficients (the

c_{ij} 's) in the objective function. In Section 3.1.1 we present a detailed outline of the reduced-size procedure, including the reformulated transportation problem. The ordering of the pairs is described in Section 3.1.2. Finally, the computation of the probabilities for the initial outcomes and the cost coefficients are given in Section 3.1.3.

3.1.1 General Outline of the Procedure

The general outline of the procedure is as follows. First, the $\binom{n}{2}$ subsets of $\{1, \dots, n\}$ of size 2 are ordered in a manner to be described later. (For now, we simply note that any ordering can be used to reduce the size of the transportation problem. The specific one used is for the purpose of accomplishing the size reduction while also attempting to give up as little as possible of the gains in overlap that the optimal procedure yields.) We let I_i , $i = 1, \dots, \binom{n}{2}$, denote the i -th element in the ordering; let $I_{\binom{n}{2}+1}, \dots, I_{\binom{n}{2}+n}$ be the n singleton subsets; and set $I_{\binom{n}{2}+n+1} = \emptyset$. Thus, the I_i 's constitute all subsets of $\{1, \dots, n\}$ of 2 or fewer elements. For each possibility for I , a unique set I^* is associated among these $\binom{n}{2} + n + 1$ subsets and the new selection probabilities conditioned on the associated I^* , rather than on I itself. Therefore, the new selection probabilities are conditioned on $\binom{n}{2} + n + 1$ events instead of a possible 2^n events, which is the reason for the size reduction. The associated I^* is the first I_i for which $I_i \subset I$. That is, if I consists of at least two integers, the associated I^* is the first pair in the ordering contained in I , while if I is a singleton set or empty then $I^* = I$.

The reduced-size transportation problem attempts to retain the PSUs corresponding to elements in the associated set I^* in the new sample, but does not use information on elements in $I \sim I^*$. The form of this reduced-sized transportation problem based on the set of I_i 's is as follows. Let p_i^* be the probability that $I^* = I_i$, $i = 1, \dots, \binom{n}{2} + n + 1$, and abbreviate $\pi_j^* = P(S_j)$, $j = 1, \dots, \binom{n}{2}$. For each i, j , the variable x_{ij} is the joint probability that $I^* = I_i$ and that $N = S_j$, while c_{ij} is the expected number of elements in $I \cap S_j$ given $I^* = I_i$. The problem to solve is to determine $x_{ij} \geq 0$ that maximize

$$\sum_{i=1}^{\binom{n}{2}+n+1} \sum_{j=1}^{\binom{n}{2}} c_{ij} x_{ij}, \quad (3.1)$$

subject to

$$\sum_{j=1}^{\binom{n}{2}} x_{ij} = p_i^*, \quad i = 1, \dots, \binom{n}{2} + n + 1, \quad (3.2)$$

$$\sum_{i=1}^{\binom{n}{2}+n+1} x_{ij} = \pi_j^*, \quad j = 1, \dots, \binom{n}{2}. \quad (3.3)$$

Once the optimal x_{ij} 's have been obtained, then the conditional new selection probabilities for $S_j, j = 1, \dots, \binom{n}{2}$, given $I^* = I_i$, are x_{ij}/p_i^* . Note that the number of variables, x_{ij} , in the formulation (3.1)–(3.3) is $(\binom{n}{2} + n + 1) \times \binom{n}{2}$, in comparison with a maximum of $2^n \times \binom{n}{2}$ in the formulation (2.1)–(2.3).

It remains to explain the general method for obtaining the ordering of the $\binom{n}{2}$ pairs and the procedures for computing the p_i^* 's and c_{ij} 's. Before doing this, we present an example of the reduced-size procedure, namely the two-PSUs-per-stratum example used in Section 2 to illustrate the transportation problem formulation for the optimal procedure.

The ordering of the pairs for this example, as will be shown later, is $\{2,3\}, \{1,2\}, \{1,3\}$. Consequently, the I_i 's, are as given in Table 3. Note that if $I = \{1,2,3\}$ or $I = \{2,3\}$, then the associated set is $I_1 = \{2,3\}$. For the other six possibilities for I the associated set is I itself.

Consequently, from Table 1 we obtain that

$$p_1^* = P(I = \{1,2,3\}) + P(I = \{2,3\}) = .525, \quad (3.4)$$

$p_i^* = P(J_i), i = 2,3$, and $p_i^* = P(J_{i+1}), i = 4, \dots, 7$, yielding the values in Table 3. Since $\pi_j^* = P(S_j)$, we have $\pi_1^* = .30, \pi_2^* = .20, \pi_3^* = .50$.

Table 3

Probabilities of Associated Sets: Reduced-Size Procedure

	<i>i</i>						
	1	2	3	4	5	6	7
I_i	$\{2,3\}$	$\{1,2\}$	$\{1,3\}$	$\{1\}$	$\{2\}$	$\{3\}$	\emptyset
p_i^*	.525	.135	.105	.045	.09	.07	.03

The c_{ij} values for this example are given in Table 4. In order to obtain these values, we simplified the computation by letting

$$b_{it} = P(t \in I \mid I^* = I_i),$$

$$i = 1, \dots, \binom{n}{2} + n + 1, \quad t = 1, \dots, n, \quad (3.5)$$

and noting that if $S_j = \{s,t\}$ then

$$c_{ij} = b_{is} + b_{it}. \quad (3.6)$$

That is, the expected number of elements in $I \cap S_j$ given $I^* = I_i$ is simply the sum of the probabilities that each of the two elements in S_j was in I given $I^* = I_i$. Also observe that while the transportation problem for the optimal procedure knows the exact value for I and hence knows with certainty whether each element in S_j was in I ,

this is not the case for the reduced-size procedure, since only the associated set I_i is known. To illustrate, consider the first row of Table 4. Since $I_1 = \{2,3\}$, we know that $2 \in I$ and $3 \in I$, and hence $b_{12} = b_{13} = 1$. However, we do not with certainty whether $1 \in I$ since I_1 is the associated set for both $I = \{1,2,3\}$ and $I = \{2,3\}$. In fact, from Table 1,

$$b_{11} = \frac{P(I = \{1,2,3\})}{P(I = \{1,2,3\}) + P(I = \{2,3\})} = .6.$$

Then $c_{11} = b_{11} + b_{12} = 1.6$, with c_{12}, c_{13} computed similarly. For the remaining six rows in Table 4, $I_i = I$ and hence it is known with certainty which integers were in I . Consequently, the c_{ij} 's for these six rows are easily computed.

Finally, we maximize the expected overlap (3.1) subject to (3.2) and (3.3), obtaining the x_{ij} values in Table 4. The conditional probabilities $P(N = S_j \mid I^* = I_i)$ in Table 5 are then obtained by dividing each of the x_{ij} entries in the i -th row of Table 4 by p_i^* .

Table 4

Values of c_{ij} and Values of x_{ij} that Maximize Overlap for the Reduced-Size Procedure

<i>i</i>	I_i	c_{ij}			x_{ij}		
		<i>j</i>			<i>j</i>		
		1	2	3	1	2	3
1	$\{2,3\}$	1.6	1.6	2.0	0.000	0.025	0.500
2	$\{1,2\}$	2.0	1.0	1.0	0.135	0.000	0.000
3	$\{1,3\}$	1.0	2.0	1.0	0.000	0.105	0.000
4	$\{1\}$	1.0	1.0	0.0	0.045	0.000	0.000
5	$\{2\}$	1.0	0.0	1.0	0.090	0.000	0.000
6	$\{3\}$	0.0	1.0	1.0	0.000	0.070	0.000
7	\emptyset	0.0	0.0	0.0	0.030	0.000	0.000

Table 5

Conditional Probabilities for the Reduced-Size Procedure

<i>i</i>	I_i	<i>j</i>		
		1	2	3
1	$\{2,3\}$	0	1/21	20/21
2	$\{1,2\}$	1	0	0
3	$\{1,3\}$	0	1	0
4	$\{1\}$	1	0	0
5	$\{2\}$	1	0	0
6	$\{3\}$	0	1	0
7	\emptyset	1	0	0

The expected overlap for the reduced-size procedure is .01 less than optimal, that is 1.725 PSUs. The deviation from optimality arises solely because the expected overlap is 1.6 for the joint event that $I^* = \{2,3\}$ and $N = \{1,3\}$. Since the probability of this joint event is .025, and the optimal procedure for this example always produces an overlap of 2 when at least 2 of the PSUs were in the initial sample, the deviation from optimality is $.025(2 - 1.6) = .01$.

The reason that the reduced-size procedure is not able to obtain optimality is that the pair $\{2,3\}$ has a smaller probability of selection in the new sample than in the initial sample. As a result, both the optimal procedure and the reduced-size procedure must sometimes select another pair (always $\{1,3\}$ for both procedures in this example) when $\{2,3\}$ was in the initial sample. The distinction between the two procedures is that the optimal procedure only selects $\{1,3\}$ when $1 \in I$. The reduced-size procedure is unable to use the information about whether $1 \in I$. As a result, when $\{2,3\} \subset I$, $1 \in N$ independently of whether $1 \in I$. This results in a deviation from the optimal overlap.

3.1.2 The Ordering of the Pairs

We now proceed to show in general how the ordering of the pairs is obtained. We use the additional notation here that p_{st} , π_{st} , $s, t = 1, \dots, n$, $s \neq t$, is the joint probability that $s, t \in I$ and $s, t \in N$, respectively.

The motivation for the ordering of the pairs is as follows. If the i -th pair in the ordering is $\{s, t\}$ then it would be possible for the transportation problem to retain this pair in the new sample when $I^* = I_i$ with conditional probability $\min\{1, \pi_{st}/p_i^*\}$. (The conditional retention probability cannot be any higher than this, since a higher value would result in an unconditional selection probability for the pair in the new design exceeding π_{st} .) Therefore, roughly the goal in the ordering is to make these conditional probabilities as large as possible on average over all pairs.

To illustrate how the ordering of the pairs affects the expected overlap we consider the example of Table 3. Our ordering procedure, as will be shown later, produces the indicated ordering and yields an expected overlap of 1.725 PSUs. Next consider the following alternative ordering for this example. Let the first pair in the ordering be $\{1,3\}$, the second pair be $\{1,2\}$ and the last pair be $\{2,3\}$. With this alternative ordering, $I^* = \{1,3\}$ whenever either $I = \{1,2,3\}$ or $I = \{1,3\}$. Therefore, for this ordering p_i^* is the probability that $I^* = \{1,3\}$, which is now .42. Furthermore, for this alternative ordering, $p_3^* = P(I^* = \{2,3\}) = P(I = \{2,3\}) = .21$, while the other 5 columns in Table 3 remain unchanged. The alternative ordering results in a table of conditional probabilities similar to Table 5, except that in row 1 the I_i , $j = 2$ and $j = 3$ columns now become $\{1,3\}$, 10/21 and 11/21, respectively, and in row 3 the corresponding columns are now $\{2,3\}$, 0 and 1, respectively.

It can be calculated, using the same approach used for the original ordering that the expected overlap for the alternative ordering is 0.055 less than optimal, that is 1.68 PSUs. The reason that this alternative ordering results in a lower expected overlap is as follows. In general a later placement of a pair in the ordering, results in a lower value for the corresponding p_i^* , and hence a higher conditional retention probability when $I^* = I_i$. That is, with $\{1,3\}$ first in the ordering, $\pi_{13}/p_1^* = 10/21$, which is the conditional retention probability for this pair when $I^* = \{1,3\}$; while when $\{1,3\}$ is third in the ordering, $\pi_{13}/p_3^* > 1$ and this pair is retained with certainty. Now the conditional retention probability for the pair $\{2,3\}$ when $I^* = \{2,3\}$ also increases to 1 when $\{2,3\}$ is moved from first to third in the ordering, but the increase is only from 20/21, and hence the original ordering in Table 3 produces a higher expected overlap than the alternative ordering.

Thus, as this example illustrates, the goal of the ordering is to place pairs earlier in the ordering that have a relatively high conditional retention probability even with an early placement. To obtain the desired ordering of the pairs of integers, an ordering $f(1), \dots, f(n)$ of $\{1, \dots, n\}$ will first be obtained by recursion. Then corresponding to each $k = 1, \dots, n - 1$, an ordering $g_k(1), \dots, g_k(n - k)$ of $\{1, \dots, n\} \sim \{f(1), \dots, f(k)\}$ will be constructed by recursion. A linear ordering of the distinct pairs in $\{1, \dots, n\}$ would then be determined as follows. Each such pair can be represented uniquely as an ordered pair $(f(k), g_k(\ell))$ for some $k \in \{1, \dots, n - 1\}$, $\ell \in \{1, \dots, n - k\}$. A second pair representable in the form $(f(k'), g_{k'}(\ell'))$ precedes $(f(k), g_k(\ell))$ if and only if either $k' < k$, or $k' = k$ and $\ell' < \ell$. To illustrate, for the example just considered it will be shown later that $f(1) = 2, f(2) = 3, f(3) = 1, g_1(1) = 3, g_1(2) = 1, g_2(1) = 1$, and hence the ordering of the pairs is $\{2,3\}, \{2,1\}, \{3,1\}$. Both the f ordering and the g_k ordering will be constructed to meet the goal stated at the beginning of this paragraph.

To obtain the ordering $f(1), \dots, f(n)$, recursively define $f(k)$, $k = 1, \dots, n$, by choosing $f(k) \in T_k$ satisfying

$$\pi_{f(k)}/p_{f(k)}^{(k)} = \max\{\pi_i/p_i^{(k)} : i \in T_k\},$$

where

$$T_1 = \{1, \dots, n\}, \quad T_k = T_{k-1} \sim \{f(k-1)\},$$

$$k = 2, \dots, n, \quad p_i^{(k)} = P(i \in I \text{ and } I \subset T_k),$$

$$k = 1, \dots, n, \quad i \in T_k. \quad (3.7)$$

Since $p_i^{(1)} = p_i$, the ordering just defined corresponds to placing first a PSU with the greatest value of π_i/p_i^* . For all k , $p_{f(k)}^{(k)}$ is the probability that $f(k)$ was in I and none of the $k - 1$ elements preceding $f(k)$ in the f ordering were in I , and hence $p_{f(k)}^{(k)}$ is the probability that

an attempt is made to retain $A_{f(k)}$ in the new sample either as the first member of an ordered pair of initial sample PSUs or as the only initial sample PSU in S . Generally, the larger $\pi_{f(k)}/p_{f(k)}^{(k)}$ is, the greater the probability that this attempt would be successful. Thus, the motivation for the f ordering of the individual PSUs is the analog of the motivation for the ordering of the pairs of PSUs that we previously discussed.

It remains to explain how to compute $p_i^{(k)}$ for $k \geq 2$. To this end, let r denote the number of initial strata with PSUs in common with S and let F_α , $\alpha = 1, \dots, r$, denote a partition of $\{1, \dots, n\}$ such that i and j are in the same F_α if and only if A_i and A_j were in the same initial stratum. Then let

$$p'_\alpha(T) = P(I \cap F_\alpha \subset T), \quad \alpha = 1, \dots, r, \\ T \subset \{1, \dots, n\}, \quad (3.8)$$

$$p''_{i\alpha}(T) = P(i \in I \text{ and } I \cap F_\alpha \subset T), \quad \alpha = 1, \dots, r, \\ T \subset \{1, \dots, n\}, \quad i \in F_\alpha \cap T, \quad (3.9)$$

and observe that

$$p'_\alpha(T) = 1 - \sum_{i \in F_\alpha \sim T} p_i + \sum_{\substack{i, j \in F_\alpha \sim T \\ i < j}} p_{ij}, \quad (3.10)$$

$$p''_{i\alpha}(T) = p_i - \sum_{j \in F_\alpha \sim T} p_{ij}, \quad (3.11)$$

and finally, as established in Ernst and Ikeda (1994),

$$p_i^{(k)} = p''_{i\alpha}(T_k) \prod_{\substack{\ell=1 \\ \ell \neq \alpha}}^r p'_\ell(T_k), \quad k = 1, \dots, n, \\ i \in F_\alpha \cap T_k. \quad (3.12)$$

Next, for each $k = 1, \dots, n-1$, the ordering $g_k(\ell)$, $\ell = 1, \dots, n-k$, is recursively defined by choosing $g_k(\ell) \in T_{k\ell}$ satisfying

$$\pi_{f(k), g_k(\ell)} / p_{f(k), g_k(\ell)}^{(\ell)} = \max\{\pi_{f(k), j} / p_{f(k), j}^{(\ell)} : j \in T_{k\ell}\},$$

where

$$T_{k1} = \{1, \dots, n\} \sim \{f(1), \dots, f(k)\}, \\ T_{k\ell} = T_{k(\ell-1)} \sim \{g_k(\ell-1)\}, \quad \ell = 2, \dots, n-k, \\ T_{k\ell}^* = T_{k\ell} \cup \{f(k)\}, \quad \ell = 1, \dots, n-k, \\ p_{f(k), j}^{(\ell)} = P(f(k), j \in I \text{ and } I \subset T_{k\ell}^*), \\ \ell = 1, \dots, n-k, \quad j \in T_{k\ell}. \quad (3.13)$$

Note that $p_{f(k), j}^{(\ell)}$ is thus the joint probability that $f(k)$ is the first integer in the f ordering in I , that none of the first $\ell-1$ integers in the g_k ordering are in I , and that $j \in I$. Consequently, $p_{f(k), g_k(\ell)}^{(\ell)}$ is the probability that $I^* = \{f(k), g_k(\ell)\}$. Furthermore, if $I_i = \{f(k), g_k(\ell)\}$ then $p_i^* = p_{f(k), g_k(\ell)}^{(\ell)}$, and hence the choice of $g_k(\ell)$ results in the largest value of $\pi_{f(k), g_k(\ell)} / p_i^*$ among the elements in $T_{k\ell}$ in accordance with the previously stated goal for the ordering of the pairs of PSUs.

To compute $p_{f(k), j}^{(\ell)}$, it is established in Ernst and Ikeda (1994) that if $f(k) \in F_\alpha$, $j \in F_\beta$, then

$$p_{f(k), j}^{(\ell)} = p_{f(k), j} \prod_{\substack{\ell=1 \\ \ell \neq \alpha}}^r p'_\ell(T_{k\ell}^*) \text{ if } \alpha = \beta, \\ = p''_{f(k), \alpha}(T_{k\ell}^*) p''_{j\beta}(T_{k\ell}^*) \prod_{\substack{\ell=1 \\ \ell \neq \alpha, \beta}}^r p'_\ell(T_{k\ell}^*) \text{ if } \alpha \neq \beta. \quad (3.14)$$

We illustrate the computations used in obtaining the ordering for the example that we have been considering. First note that $f(1) = 2$ since the largest value of π_i / p_i occurs for $i = 2$. Next we find $g_1(1)$ which, since $f(1) = 2$, is the $j \in \{1, 3\}$ with the maximum value of $\pi_{2j} / p_{2j}^{(1)}$. To find this j , first let $F_\alpha = \{\alpha\}$, $\alpha = 1, 2, 3$, and note that $T_{11}^* = \{1, 2, 3\}$. From (3.14) with $\alpha = 2$, $\beta = 1$, it then follows that

$$p_{21}^{(1)} = p''_{22}\{1, 2, 3\} p''_{11}\{1, 2, 3\} p'_3\{1, 2, 3\} = p_2 p_1 \cdot 1 = .45,$$

and similarly it can be obtained that $p_{23}^{(1)} = .525$. Hence $g_1(1) = 3$, since $.5/.525 > .3/.45$. Therefore, the first pair in the ordering is $\{f(1), g_1(1)\} = \{2, 3\}$. Then $g_1(2) = 1$, since 1 is the only integer remaining to be used in the g_1 ordering, and consequently the second pair in the ordering is $\{f(1), g_1(2)\} = \{2, 1\}$. It is not really necessary to determine $f(2)$, since $\{1, 3\}$ is the only remaining pair, and hence the last pair, but to further illustrate the computations, observe that $T_2 = \{1, 3\}$, $p_1^{(2)} = p''_{11}\{1, 3\} p''_{22}\{1, 3\} p'_3\{1, 3\} = p_1(1 - p_2) \cdot 1 = .15$ by (3.12), and similarly $p_3^{(2)} = p_3(1 - p_2) \cdot 1 = .175$. Hence $f(2) = 3$, since $.7/.175 > .5/.15$. Consequently, $g_2(1) = 1, f(3) = 1$.

3.1.3 Computation of p_i^* and c_{ij}

Next we explain the computation of the p_i^* 's. If I_i consists of the pair of integers $I_i = \{f(k), g_k(\ell)\}$ then, as previously noted, $p_i^* = p_{f(k), g_k(\ell)}^{(\ell)}$. Consequently, p_i^* can be computed from (3.14) with $j = g_k(\ell)$.

If I_i is a singleton set $\{t\}$ for some $t \in F_\alpha$, then, as established in Ernst and Ikeda (1994),

$$p_i^* = p_{i\alpha}''(\{t\}) \prod_{\substack{u=1 \\ u \neq \alpha}}^r p_u'(\emptyset). \quad (3.15)$$

Finally, if $I_i = \emptyset$, then

$$p_i^* = \prod_{u=1}^r p_u'(\emptyset).$$

It remains only to explain how to compute the c_{ij} 's which, by (3.5) and (3.6), reduces to computing b_{it} , $i = 1, \dots, \binom{n}{2} + n + 1$, $t = 1, \dots, n$.

To compute b_{it} , observe that

$$\begin{aligned} b_{it} &= 0 \quad \text{if } I_i = \emptyset, \\ &= 1 \quad \text{if } I_i = \{v\} \quad \text{and } t = v, \\ &= 0 \quad \text{if } I_i = \{v\} \quad \text{and } t \neq v, \end{aligned}$$

while if $I_i = \{f(k), g_k(\ell)\}$ and $f(k) \in F_\alpha$, $g_k(\ell) \in F_\beta$, $t \in F_\gamma$, then

$$b_{it} = 1 \quad \text{if } t = f(k) \quad \text{or } t = g_k(\ell), \quad (3.16)$$

$$= 0 \quad \text{if } t \notin T_{k\ell}^*, \quad (3.17)$$

$$\begin{aligned} &= 0 \quad \text{if } t \in T_{k\ell} \sim \{g_k(\ell)\} \\ &\quad \text{and } \gamma = \alpha = \beta, \end{aligned} \quad (3.18)$$

$$\begin{aligned} &= \frac{p_{f(k),t}}{p_{f(k),\alpha}''(T_{k\ell}^*)} \quad \text{if } t \in T_{k\ell} \sim \{g_k(\ell)\} \\ &\quad \text{and } \gamma = \alpha \neq \beta, \end{aligned} \quad (3.19)$$

$$\begin{aligned} &= \frac{p_{g_k(\ell),t}}{p_{g_k(\ell),\beta}''(T_{k\ell}^*)} \quad \text{if } t \in T_{k\ell} \sim \{g_k(\ell)\} \\ &\quad \text{and } \gamma = \beta \neq \alpha, \end{aligned} \quad (3.20)$$

$$\begin{aligned} &= \frac{p_{t\gamma}''(T_{k\ell}^*)}{p_{t\gamma}'(T_{k\ell}^*)} \quad \text{if } t \in T_{k\ell} \sim \{g_k(\ell)\} \\ &\quad \text{and } \gamma \neq \alpha, \gamma \neq \beta. \end{aligned} \quad (3.21)$$

In Ernst and Ikeda (1994) it is demonstrated how (3.16)–(3.21) were obtained.

In the actual implementation for the SIPP application, modifications of the reduced-size procedure were needed to overlap the 1990s SIPP design with the 1980s SIPP design. The modifications were necessary because the PSU definitions in the 1980s and 1990s designs were not identical. As a result, some PSUs in the 1990s design could intersect more than one 1980s design PSU. These modifications are detailed in Ernst and Ikeda (1994).

3.2 Modifications of Reduced-Sized Procedure for Other Designs

In general, consider any m' -PSUs-per-stratum without replacement initial design and any m -PSUs-per-stratum without replacement final design, where m' , m are any positive integers. Although the reduced-size procedure in Section 3.1 was only presented for the case $m = m' = 2$, it is actually applicable for any m, m' . We will sketch the modifications necessary when $m \neq 2$ or $m' \neq 2$.

A different value of m' only requires modification of some of the computations. For example, if $m = 2$, but $m' \neq 2$, then the computations for $p_i^{(k)}$, $p_{f(k),j}^{(\ell)}$ and c_{ij} would be different but their definitions would not change.

If $m = 3$, then, regardless of the value of m' , the set of all distinct triples, instead of pairs, of integers in $\{1, \dots, n\}$, is ordered. If I consists of at least three integers, then the new selection probabilities are conditioned only on the first listed triple in the ordering contained in I . Otherwise, the new selection probabilities are conditioned on I itself. Thus the new selection probabilities are conditioned on $\binom{n}{3} + \binom{n}{2} + n + 1$ events.

To obtain the desired ordering of the triples of integers, first the orderings $f(1), \dots, f(n)$ and $g_k(1), \dots, g_k(n - k)$ are constructed exactly as in the case $m = 2$. Then, corresponding to each $k = 1, \dots, n - 2$, $\ell = 1, \dots, n - k - 1$, an ordering $h_{k\ell}(1), \dots, h_{k\ell}(n - k - \ell)$ of $\{1, \dots, n\} \sim \{f(1), \dots, f(k), g_k(1), \dots, g_k(\ell)\}$ is constructed in a manner similar to the construction of $g_k(1), \dots, g_k(n - k)$. For example, in defining $h_{k\ell}(v)$ for $v \geq 2$, $p_{f(k),j}^{(\ell)}$ in the definition of $g_k(\ell)$ is replaced by

$$P(f(k), g_k(\ell), j \in I \quad \text{and} \quad I \subset (T_{k\ell}^* \cup g_k(\ell)) \sim \{h_{k\ell}(1), \dots, h_{k\ell}(v - 1)\}).$$

A linear ordering of the distinct triples in $\{1, \dots, n\}$ is then determined by representing each triple uniquely as an ordered triple of the form $(f(k), g_k(\ell), h_{k\ell}(v))$. A second triple $(f(k'), g_{k'}(\ell'), h_{k'\ell'}(v'))$ precedes the first if and only if either $k' < k$, or $k' = k$ and $\ell' < \ell$, or $k' = k$ and $\ell' = \ell$ and $v' < v$.

For $m \geq 4$, ordered m -tuples would be defined in a similar manner and the new selection probabilities conditioned on $\binom{n}{m} + \binom{n}{m-1} \dots + n + 1$ events.

For $m = 1$, the new selection probabilities are conditioned on the first member of the ordering $f(1), \dots, f(n)$ in I if $I \neq \emptyset$, or on \emptyset if $I = \emptyset$.

Note that if $m > m'$, it is possible that at least some ordered m -tuples cannot be subsets of I , in which case all such subsets should be excluded from the ordering and the set of events on which the new selection probabilities are conditioned. If no m -tuple can be a subset of I , then the new selection probabilities are conditioned on I itself.

It is not necessary to limit the initial events used in the transportation problem to subsets of I of size m or less. For example, if $m = 2$ and $\binom{n}{3} + \binom{n}{2} + n + 1$ is sufficiently small, then a procedure conditioned on subsets of three or less can be used, resulting in a generally higher expected overlap. Conversely, if $\binom{n}{m} + \binom{n}{m-1} \dots + n + 1$ is too large, the new selection probabilities can be conditioned on subsets of I of size m'' or less, where $m'' < m$, although with a generally smaller expected overlap.

3.3 Relationship Between Expected Overlap for the Reduced-Size Procedure, the Optimal Procedure and Independent Selection

Let Ω_I , Ω_R , Ω_O denote the expected overlap for the independent selection, the reduced-size procedure, and the optimal procedure, respectively. In Ernst and Ikeda (1994) the relationship between these quantities is explored. We briefly summarize here some of the results.

It is established that $\Omega_I \leq \Omega_R \leq \Omega_O$ for any m, m' where m, m' are as in Section 3.2. In addition, for the case that we have been focusing on, $m = m' = 2$, lower bounds are established on Ω_R and upper bounds are established on Ω_O and $\Omega_O - \Omega_R$.

For example, let μ_2 denote the probability that there are at least two elements in I , μ_1 denote the probability that I is a singleton set, and

$$\lambda = \min\{\min\{\pi_i/p_i: i = 1, \dots, n\}, \min\{\pi_{ij}/p_{ij}: i, j = 1, \dots, n, i \neq j\}, 1\}.$$

Then $\Omega_O \leq 2\mu_2 + \mu_1$, $\Omega_R \geq \lambda(2\mu_2 + \mu_1/2)$, and $\Omega_O - \Omega_R \leq 2(1 - \lambda)\mu_2 + (1 - \lambda/2)\mu_1$.

Unfortunately these bounds are not always very tight. However, in certain circumstances they are useful. For example, if $\pi_{ij} \geq p_{ij}$ for all i, j and the probability is 1 that there is at least two elements in I , then it follows from these bounds that $\Omega_R = \Omega_O = 2$.

Finally, an example is presented to illustrate a worst case situation for Ω_R in relation to Ω_O for the case $m, m' = 2$. It shows that Ω_O may equal 2, while Ω_R is arbitrarily close to 0. Thus, at least in theory, the reduced-size procedure can be ineffective. However, in practice, as will be shown in the next section, Ω_R is much closer to Ω_O than to Ω_I , at least for the SIPP application.

4. APPLICATION OF REDUCED-SIZE PROCEDURE TO SIPP

Results from simulations of the SIPP overlap, done prior to production for research and testing purposes, are presented, as well as results from the actual SIPP production overlap. Further details are given in Ernst and Ikeda (1992b, 1994).

In the implementation of the reduced-size overlap procedure, minimum cost flow (MCF) optimization software, written by Darwin Kingman and John Mote at the University of Texas at Austin, was used to solve the required transportation problem. A FORTRAN program was written to produce input to and process output from the MCF software.

To test the software prior to production, the program was used to overlap two stratifications, based on 1970 census data, of the SIPP Midwest region with the actual 1980s design stratification for the SIPP Midwest region. (At the time of this test, 1990 census data was not yet available.) The 1970-based stratifications were produced by stratifying the 1980s SIPP noncertainty PSUs in the Midwest region using 1970 data. Both of the 1970-based stratifications partitioned the noncertainty PSUs into 31 strata, using different sets of stratification variables. The stratifications based on 1980 and 1970 data were treated as "initial" and "final" stratifications for the purposes of the overlap algorithm.

In the actual implementation, as noted in Section 3.1 and detailed in Ernst and Ikeda (1994), a modification of the reduced-size procedure was used to overlap the 1990s SIPP design with the 1980s SIPP design, because the PSU definitions in the 1980s and 1990s designs were not identical. The modified reduced-size procedure was used to overlap 103 final (1990s design) nonselfrepresenting strata in SIPP.

The expected overlap was calculated for the reduced-size maximum overlap algorithm, for independent selection of final PSUs, and for an upper bound to the expected overlap for the optimal procedure. An upper bound was calculated instead of the actual optimal overlap, since the optimal overlap cannot be calculated for the larger strata. For the simulation, the upper bound used is the one stated in Section 3.3, $\mu_2 + 2\mu_1$, while for the production SIPP, a different upper bound, described in Ernst and Ikeda (1994), was required because the PSU definitions in the 1980s and 1990s were not identical.

The results from the two final stratifications in the simulation were generally similar to each other. Combining the results from both stratifications, the mean expected overlap for this set of 62 strata was 1.552, 1.569 and 0.480 PSUs/stratum for the reduced-size procedure, the upper bound to the optimal overlap and independent selection respectively. For the actual SIPP implementation, the corresponding number was 1.523, 1.647 and 0.582, respectively, while the corresponding expected number of PSUs overlapped for the 103 strata was 156.9, 169.6 and 59.9, respectively. Thus, in both the simulations and the production SIPP, the reduced-size procedure yielded results reasonably close to the upper bound for the optimal procedure.

The reduced-size algorithm took a fairly short time to run on most strata. The CPU times in the simulation for

final strata with different numbers of PSUs are given below. The reduced-size program was run on a Solbourne 5/605 computer. The median number of PSUs in a stratum, for the entire group of 62 strata, was 17 PSUs. The 68 PSUs stratum was the largest stratum.

Table 6

CPU Times for Reduced-Size Procedure

Number of PSUs	CPU Time (hrs:min:sec)
18	0:36
37	5:44
49	24:05
68	2:23:43

We also calculated for the actual SIPP implementation, that of the 103 final strata overlapped by the modified reduced-size procedure, 41 would not have run under the optimal procedure. This calculation was based on our estimate that the maximum size transportation problem, in terms of number of variables, that could have run in production was 4×10^6 . The number of variables for the optimal procedure was less than 4×10^6 for all 56 strata for which $n \leq 14$, but exceeded this limit for all but 6 of the 47 strata with $n \geq 15$, including two with $n = 15$. The maximal size of the transportation for the optimal procedure among the 103 strata occurred for a stratum with $n = 46$, for which there were 3.61×10^{12} variables. In contrast, there were 1.03×10^6 variables for the modified reduced-size procedure for this stratum.

Another question of interest is the overlap effectiveness of the reduced-size procedure in comparison with the overlap procedure of Ernst (1986). In general it is believed that the reduced-size procedure should produce a higher overlap in situations when both are usable, since the reduced-size procedure makes use of the stratum-to-stratum independence in the initial design. However, although the procedure in Ernst (1986) is applicable to two-PSU-per-stratum designs, no computer program has ever been written at the Census Bureau (or anywhere else that the authors are aware of) to implement this procedure for such designs, since there has not yet been a production application for this program. Consequently, we cannot make a direct comparison of these two methods on the same data. However, a crude comparison can be made from the results of the reduced-size overlap procedure for SIPP data and the results of the overlap using the procedure in Ernst (1986) for the overlap of 1990s CPS and NCVS designs with their respective 1980s designs. (Both the 1980s and 1990s designs for CPS and NCVS are one-PSU-per-stratum designs.)

For CPS, the overlap procedure resulted in an average increase in expected overlap, in comparison with independent selection, of .26 PSUs/stratum, and for NCVS the overlap procedure resulted in an average increase in expected overlap of .30 PSUs/stratum. This compares with an increase of .94 PSUs/stratum for the reduced-size procedure over independent selection for SIPP. If the two overlap procedures are equally effective, then one might expect that the increase in overlap per stratum for SIPP would be roughly twice as large as for CPS and NCVS, since SIPP has a two-PSUs-per-stratum design. By this standard, the reduced-size procedure program performs better than the procedure in Ernst (1986). However, since the stratifications were quite different for these three surveys, the validity of this comparison is open to question.

For the example considered in Sections 2 and 3, a valid comparison of the different overlap procedures can be made, since the expected overlap values for the procedure in Ernst (1986)), 1.625, was easily calculated by hand. For the reduced-size procedure the corresponding overlap value is 1.725, and for the optimal procedure it is 1.735.

CONCLUSIONS

The reduced-size overlap procedure presented in this paper meets its two key objectives in practice. It reduces the size of the transportation problems to a usable size, as evidenced both by the size of the transportation problem in the formulation (3.1)–(3.3), and the fact that it has actually been implemented in the redesign of a major survey. In addition, the procedure accomplishes the size reduction while yielding nearly optimal overlap, at least for the SIPP application. It can only be used when the PSUs in the initial design are selected independently from stratum to stratum, but when this condition is met we believe it is the overlap procedure of choice for large strata.

ACKNOWLEDGEMENTS

The programming assistance of Todd Williams is gratefully acknowledged. The authors would also like to thank the referees and the editors for their constructive comments. The views expressed in this paper are attributable to the authors and do not necessarily reflect those of the Bureau of Labor Statistics and the Census Bureau.

REFERENCES

ARAGON, J., and PATHAK, P.K. (1990). An algorithm for optimal integration of two surveys. *Sankhyā: The Indian Journal of Statistics*, 52, 198-203.

ARTHANARI, T.S., and DODGE, Y. (1981). *Mathematical Programming in Statistics*. New York: John Wiley and Sons.

- CAUSEY, B.D., COX, L.H., and ERNST, L.R. (1985). Applications of transportation theory to statistical problems. *Journal of the American Statistical Association*, 80, 903-909.
- ERNST, L.R. (1986). Maximizing the overlap between surveys when information is incomplete. *European Journal of Operational Research*, 27, 192-200.
- ERNST, L.R. (1989). Further Applications of Linear Programming to Sampling Problems. Bureau of the Census, Statistical Research Division, Research Report Series, No. RR-89/05.
- ERNST, L.R., and IKEDA, M. (1992a). Modification of the Reduced-Size Transportation Problem for Maximizing Overlap When Primary Sampling Units Are Redefined in the New Design. Bureau of the Census, Statistical Research Division, Technical Note Series, No. TN-91/01.
- ERNST, L.R., and IKEDA, M. (1992b). Summary of the Performance of the Maximum Overlap Algorithms for the 1990's Redesign of the Demographic Surveys. Bureau of the Census, Statistical Research Division, Technical Note Series, No. TN-92/01.
- ERNST, L.R., and IKEDA, M. (1994). A Reduced-Size Transportation Algorithm for Maximizing the Overlap Between Surveys. Bureau of the Census, Statistical Research Division, Research Report Series, No. RR-93/02.
- GLOVER, F., KARNEY, D., KLINGMAN, D., and NAPIER, A. (1974). A computation study on start procedures, basic change criteria and solution algorithms for transportation problems. *Management Sciences*, 20, 793-813.
- KEYFITZ, N. (1951). Sampling with probabilities proportional to size: Adjustment for changes in probabilities. *Journal of the American Statistical Association*, 46, 105-109.
- KISH, L., and SCOTT, A. (1971). Retaining units after changing strata and probabilities. *Journal of the American Statistical Association*, 66, 461-470.
- PATHAK, P.K., and FAHIMI, M. (1992). Optimal integration of surveys. In *Essays in Honor of D. Basu*. Eds. M. Ghosh, and P.K. Pathak. Hayward, California: Institute of Mathematical Statistics, 208-224.
- PERKINS, W.M. (1970). 1970 CPS Redesign: Proposed Method for Deriving Sample PSU Selection Probabilities Within 1970 NSR Stata. Memorandum to Joseph Waksberg, Bureau of the Census.
- RAJ, D. (1968). *Sampling Theory*. New York: McGraw Hill.

How Prenotice Letters, Stamped Return Envelopes and Reminder Postcards Affect Mailback Response Rates for Census Questionnaires

DON A. DILLMAN, JON R. CLARK and MICHAEL D. SINCLAIR¹

ABSTRACT

In a 1992 National Test Census the mailing sequence of a prenotice letter, census form, reminder postcard, and replacement census form resulted in an overall mailback response of 63.4 percent. The response was substantially higher than the 49.2 percent response rate obtained in the 1986 National Content Test Census, which also utilized a replacement form mailing. Much of this difference appeared to be the result of the prenotice – census form – reminder sequence, but the extent to which each main effect and interactions contributed to overall response was not known. This paper reports results from the 1992 Census Implementation Test, a test of the individual and combined effectiveness of a prenotice letter, a stamped return envelope and a reminder postcard, on response rates. This was a national sample of households ($n = 50,000$) conducted in the fall of 1992. A factorial design was used to test all eight possible combinations of the main effects and interactions. Logistic regression and multiple comparisons were employed to analyze test results.

KEY WORDS: Mail survey; Response rates; Multiple comparisons; Logistic regression.

1. INTRODUCTION

A decline of 10 percentage points from 75 to 65 in the mailback response rates for the 1990 U.S. Decennial Census has stimulated the conduct of research aimed at finding ways to improve response. Each percentage point gain in response has the potential for saving approximately \$16 million in personal visit enumeration costs (Miskura 1992). From an earlier experiment it was learned that respondent-friendly construction and asking somewhat fewer questions than posed in the 1990 Census short questionnaire improved mailback response rates by 8.0 percentage points (Dillman, Clark and Sinclair 1993). An experimental census form with these features was returned by 71.4 percent of households, compared to 63.4 percent of those which had received the 1990 Census short form as a control. Response rates for both of these forms were substantially higher than had previously been obtained in similar non-census year tests. For example, in the 1986 National Content Test which utilized a questionnaire equivalent to the 1990 Census short form, a 49.2 response rate was obtained. It was hypothesized that part of the high response observed in the recent experiment was due to a multiple contact implementation strategy which consisted of a prenotice letter, a reminder postcard and a replacement questionnaire.

The purpose of this paper is to report results of the 1992 Implementation Test (IT), a test designed to determine the relative and combined contribution to mailback response of the prenotice letter and reminder postcard used in the

previously reported experiment (Dillman *et al.* 1993). Also included in the test is the effect of including a stamped return envelope (vs. business reply) with the mailed census form.

The 1990 U.S. Decennial Census required surveying over 100,000,000 households. Cost considerations alone suggest the importance of learning the extent to which each of these three response-inducing techniques might be employed in improving household response. Although past research has suggested that each of the three elements can be important to improving response, little information is available on potential interactions among them. The study was designed in such a way as to explore the extent to which their combined uses are additive and/or interactive.

1.1 Past Research

Numerous studies have confirmed that the most important determinant of overall response to mail surveys is the number of contacts (*e.g.*, Scott 1961 and Heberlein and Baumgartner 1978). Both prenotices and reminders have been demonstrated as being effective promoters of response (*e.g.*, Kanuk and Berenson 1975, Linsky 1975 and Fox *et al.* 1988). However, past research has provided minimal insight into their relative importance as inducers of response.

Past research is generally consistent in suggesting that inclusion of a stamped return envelope (vs. a business reply envelope) improves response (Scott 1961, Kanuk and Berenson 1975, Duncan 1979, Harvey 1987 and Fox *et al.* 1988). A noteworthy exception is a regression analysis of previous studies by Heberlein and Baumgartner, which

¹ Don A. Dillman, Washington State University, Pullman, WA, U.S.A.; Jon R. Clark, U.S. Bureau of the Census, Washington DC 20233, U.S.A.; and Michael D. Sinclair, Response Analysis Corp., Princeton, NJ, U.S.A.

found no significant effect for the inclusion of stamped return envelopes (1979). A review study by Armstrong and Luske reported 20 studies in which alternatives to business reply envelopes had been tested (1987). In each of these comparisons the absolute level of response to the alternative was significantly higher in 15 of the 20 cases, by an average of 9.2 percentage points. Six studies of metered marks vs. envelopes with real stamps were reported. On average they showed a 3.4 percentage point advantage for stamps. Finally, four studies in which a constellation of response inducing factors was used to insure high overall response rates showed a 2-4 percentage point advantage for stamped over business reply envelopes (Dillman 1978).

The three response stimuli to be tested here are among the top eight techniques reported consistently in the research literature as factors which improve mailback response rates. Others include financial incentives, special postage, choice of sponsor, personalization and interest (or salience) (Dillman 1991).

Two of these eight factors, financial incentives and special postage (*e.g.*, certified or two day priority mail) were judged impractical for use in a census of more than 100,000,000 households. A third factor, sponsorship by the U.S. Bureau of the Census, was considered desirable from the standpoint of encouraging response. A fourth factor, respondent interest, or question salience could not be manipulated in the sense that the survey questions are specified by federal laws. The fifth factor, personalization of correspondence was limited by the fact that Census forms cannot be addressed to individuals and are necessarily sent to only household addresses. By examining the individual and combined response effects of the prenotice, stamped return envelope and reminder, we hoped to learn whether the use of one or more of these elements would substitute for another, therefore making it possible to improve response at less cost.

1.2 Design and Integration of Treatment Elements

Certain features of the census form mailout packet suggest that it may be overlooked or ignored by those to whom it is sent. By necessity it is sent only to household addresses; names cannot be used to address any of the letters. Accurate processing of returned questionnaires requires identification of the household address on the questionnaire itself. Separately addressing an outside envelope, letter and questionnaire and being sure that the correct components are inserted into the appropriate envelope presents a serious quality control problem in a large census. Therefore it is considered important to print addresses only on one of the pieces that has to be merged together for the mailout package. Consequently, a windowed envelope through which the address on the questionnaire can be seen is used to deliver it.

The combined effect of the inability to use resident names plus size and outward appearance of the windowed envelope suggest that it contains unimportant material or perhaps, "junk mail." Also, research on nonresponse to the 1990 Census revealed that some people did not recall receiving their census questionnaire in the mail, or saw it, but did not open it, both of which might have resulted from a mass mailing appearance (Kulka *et al.* 1991).

In this experimental test the prenotice letter and reminder postcard were designed to bring attention to the envelope containing the census form. This was accomplished in five ways. First, the prenotice was developed as a letter, and the reminder as a postcard. It was reasoned that people were more likely to look at two pieces of mail which appeared different from one another. The letter format was chosen for the prenotice in order to save the more convenient postcard format for the reminder.

Second, the prenotice letter consisted of a letter from the Director of the Census Bureau with the notation "To the residents at" and the address imaged onto stationery in the normal inside address position. Our goal was to communicate that the census questionnaire which would soon arrive was specifically for people at that address. This address also doubled as an outside address, being visible through a windowed envelope, thereby avoiding the quality control concern noted for the census form mailing of merging separately addressed components.

Third, the prenotice was scheduled to be delivered a few days before the envelope containing the census form itself, and the reminder was scheduled to arrive just a few days afterwards. The mailout dates were September 21st, 24th, and 29th, respectively. It was reasoned that to be effective, a reminder (without a replacement questionnaire) should arrive within a few days of the questionnaire, before normal household cleaning would have resulted in unopened mail being thrown out.

Fourth, the wording of the prenotice, "Within the next few days you should receive. . ." and the reminder, "A few days ago you should have received. . ." were designed to encourage recipients to look for the census form. Fifth, the use of the Director's letterhead stationery and white postcard stock which showed the seal of the Department of Commerce above the reminder message, were aimed at communicating that the census questionnaire was from the government and not from some other group attempting to emulate a governmental appearance, as is sometimes done by noting, *e.g.*, "this is your official notice."

The stamped return envelope's positive influence, if any, on response may result from encouraging trust that the request is legitimate and important (otherwise why would the sender "waste" a stamp, which could be torn off and used for another purpose and/or a recipient's reluctance to throw away something of value, *i.e.*, an uncanceled stamp). The prenotice, and to some extent the reminder, could enhance the stamp's effect by getting the

envelope containing the census form opened. Also, once opened, the awareness of an uncanceled stamp could discourage throwing away the contents so that the effect of the reminder is enhanced.

In order for the prenotice, stamped return and reminder to mutually support one another, it was deemed important that first class mail be used. Had bulk rate been used, and the mailings been closely spaced, it was likely that in some households a later mailing would have arrived before an earlier one.

In sum, this test involved more than simply juxtaposing three separate test elements from the literature. The elements were operationalized in ways that improved the likelihood that each would augment effects of the others, and be feasible for use in large scale mailings. Practically, we hoped to learn whether one or more of the elements might be eliminated without a significant loss of response, thus showing how to save costs for a census mailing.

2. EXPERIMENTAL DESIGN

A factorial design, consisting of all eight of the possible combinations of the three main effects, was used for the experiment. The treatments were as follows:

1. None (control),
2. Prenotice letter only,
3. Stamped return envelope only,
4. Reminder postcard only,
5. Letter plus stamped return,
6. Stamped return plus reminder,
7. Letter plus reminder, and
8. Letter plus stamped return plus reminder.

2.1 Sample Design

The sampling universe consisted of all housing units in the questionnaire mailback areas identified by Census Bureau address files. The 449 district office (DO) areas for

the 1990 Census were selected as the geographic units for defining the strata for the test. Two strata were defined. Due to the high correlation between the minority rate (minority is defined as including all Black and Hispanic classifications) and the 1990 Census mail response rate, the stratification objectives were met by ranking the DOs by their percent minority. DOs with a combination of high minority (Black and/or Hispanic origin) population and low 1990 questionnaire mail response rates were defined as "low response areas" (LRA) and made up the first stratum. The remaining DOs were classified as "high response areas" (HRA) and constituted the second stratum.

The first stratum, consisting of 67 DOs, had a combined minority population of about 64 percent and encompassed about 11 percent of all housing units in the census mailback areas. The second stratum of 382 DOs had a combined minority population of about 15 percent. The HRA stratum had a cumulative mail response rate in the 1990 Census of approximately 10 percentage points higher than the LRA stratum.

A sample of 50,000 housing units was selected with 25,000 units in each stratum. The LRA stratum was over-sampled to concurrently study factors related to differential undercount, which falls outside the scope of this paper. Each stratum was divided into eight equally sized panels to test the eight different treatments. A systematic sample of 3,125 housing units was selected from each panel/stratum combination. Once a housing unit was selected, the seven subsequent units were also selected. The resulting households in each of the eight unit clusters were randomly allocated to a panel. Hence, all eight neighbors got different treatments. The sample was clustered to reduce the sampling variance in the panel-to-panel comparison.

The sample size selected for this study was developed by extensive data simulations which indicated that the 50,000 unit sample would be sufficient for detecting a minimum of a 3 percent difference in all pairwise treatment comparisons.

Table 1
Implementation Test Final Rates National and Stratum Level Estimates

Treatment	Response Rate (%) Estimates and Standard Errors (%)					
	National		1990 High Response Areas		1990 Low Response Areas	
	Estimate	Standard Error	Estimate	Standard Error	Estimate	Standard Error
1. Control	50.0	0.8	51.9	0.9	36.3	0.9
2. Prenotice Letter Only	56.4	0.8	58.6	0.9	40.5	0.9
3. Stamped Return Envelope Only	52.6	0.8	54.5	0.9	37.9	0.9
4. Reminder Card Only	58.0	0.8	60.2	0.9	42.0	0.9
5. Letter and Stamp	59.8	0.8	62.1	0.9	43.0	0.9
6. Stamp and Reminder	59.5	0.8	61.8	0.9	42.6	0.9
7. Letter and Reminder	62.7	0.8	65.0	0.9	45.4	0.9
8. Letter, Stamp and Reminder	64.3	0.8	66.5	0.9	47.8	0.9

3. FINDINGS

The major results from this study are presented through two analytical methods, first through multiple pairwise comparisons of treatment means and secondly through logistic regression. See Appendix for estimation procedures. Both methods provide consistent results. The overall response rates and standard errors for each of the treatments at the national and stratum levels are presented in Table 1. They range from 50.0 percent for the control group to 64.3 percent when all three main effects are applied together.

3.1 Multiple Comparisons of Mail Response Rates

Twenty eight comparisons are presented in Table 2 corresponding to all possible pairwise comparisons of the 8 treatments. Given the space restrictions in the table, the

following abbreviations were used: C = control, L = pre-notice letter, S = stamped return envelope, R = reminder postcard.

The first three comparisons in Table 2 illustrate the improvements in response that main effect components added to response individually above and beyond the control treatment. The estimated improvement in response due to the prenotice letter was 4.2 percent in the LRA stratum, 6.7 percent in the HRA stratum and 6.4 percent at the national level. The estimated improvement due to the reminder card was 5.7 percent in the LRA stratum, 8.3 percent in the HRA stratum and 8.0 percent at the national level. All of these improvements are significant. Thus, the principal finding of this study is that both the prenotice letter and the reminder card increased mail response at the national and stratum level. No significant improvements were noted for the stamped return envelope at the national or stratum level.

Table 2
Differences in Response Rates – Each Component in the Presence of Another Component

Experimental Comparisons	Response Rate Differences (%) and 90% Confidence Intervals (C.I.)					
	National		1990 Low Response Areas (LRA)		1990 High Response Areas (HRA)	
	Difference	90% C.I.	Difference	90% C.I.	Difference	90% C.I.
1. L – C	6.4	3.3 to 9.5*	4.2	0.9 to 7.5*	6.7	3.2 to 10.2*
2. S – C	2.5	–0.5 to 5.6	1.7	–1.7 to 5.0	2.7	–0.8 to 6.1
3. R – C	8.0	4.9 to 11.1*	5.7	2.4 to 9.1*	8.3	4.9 to 11.7*
4. LS – C	9.8	6.7 to 12.9*	6.8	3.4 to 10.1*	10.2	6.7 to 13.7*
5. SR – C	9.5	6.4 to 12.5*	6.4	3.0 to 9.7*	9.9	6.5 to 13.3*
6. LR – C	12.7	9.6 to 15.7*	9.2	5.8 to 12.5*	13.2	9.7 to 16.6*
7. LSR – C	14.2	11.2 to 17.2*	11.5	8.2 to 14.8*	14.6	11.3 to 18.0*
8. L – S	3.8	0.8 to 6.9*	2.5	–0.9 to 5.9	4.1	0.6 to 7.5*
9. R – L	1.6	–1.5 to 4.8	1.5	–1.9 to 5.0	1.6	–1.96 to 5.10
10. R – S	5.5	2.4 to 8.5*	4.1	0.7 to 7.5*	5.6	2.2 to 9.0*
11. LS – L	3.4	0.3 to 6.5*	2.6	–0.9 to 6.0	3.5	0.03 to 7.0*
12. SR – L	3.1	0.03 to 6.2*	2.2	–1.3 to 5.6	3.2	–0.3 to 6.6
13. LR – L	6.3	3.2 to 9.3*	5.0	1.5 to 8.4*	6.4	3.0 to 9.9*
14. LS – S	7.3	4.2 to 10.3*	5.1	1.7 to 8.5*	7.6	4.1 to 11.0*
15. SR – S	6.9	3.8 to 10.1*	4.7	1.2 to 8.2*	7.2	3.8 to 10.7*
16. LR – S	10.1	7.1 to 13.2*	7.5	4.1 to 11.0*	10.5	7.0 to 13.9*
17. LS – R	1.8	–1.3 to 4.9	1.1	–2.4 to 4.5	1.9	–1.6 to 5.4
18. SR – R	1.5	–1.6 to 4.5	0.7	–2.8 to 4.1	1.6	–1.8 to 5.0
19. LR – R	4.7	1.6 to 7.7*	3.5	–0.02 to 6.9	4.9	1.5 to 8.3*
20. LSR – L	7.9	4.8 to 10.9*	7.3	3.9 to 10.7*	7.9	4.5 to 11.4*
21. LSR – S	11.7	8.7 to 14.7*	9.8	6.4 to 13.3*	12.0	8.6 to 15.4*
22. LSR – R	6.2	3.2 to 9.3*	5.8	2.3 to 9.3*	6.3	2.9 to 9.7*
23. LSR – LS	4.4	1.4 to 7.5*	4.7	1.2 to 8.2*	4.4	1.0 to 7.8*
24. LSR – SR	4.8	1.7 to 7.8*	5.1	1.7 to 8.6*	4.7	1.3 to 8.2*
25. LSR – LR	1.6	–1.4 to 4.5	2.3	–1.1 to 5.8	1.5	–1.8 to 4.8
26. SR – LS	–0.3	–3.3 to 2.7	–0.4	–3.8 to 3.1	–0.3	–3.7 to 3.1
27. LR – LS	2.9	–0.2 to 6.0	2.4	–1.1 to 5.9	2.9	–0.6 to 6.4
28. LR – SR	3.2	0.2 to 6.2*	2.8	–0.6 to 6.2	3.3	–0.1 to 6.6

A C.I. marked with an * indicates the difference was statistically significant at $\alpha = .10$ (9-in-10 chance that the C.I.s will include the actual differences).

3.2 Logistic Regression Analysis

A model including components for the stratum, pre-notice letter, stamp and reminder card including all of the interaction terms was evaluated. Modeling was also performed at the stratum level using only parameters for the component effects and their interactions.

The results of the full model analysis indicate that only the main effects of the letter and the reminder card along with the intercept and stratum term are statistically significant in the model. Given these results, additional modeling at the national level was accomplished with a reduced model including only the stratum main effect, the individual components and the component interactions. The results of this modeling are presented in Table 3 below.

Table 3

Analysis of Weighted Least Squares Logistic Regression
Modeling Reduced Model, no Stratum by
Component Interactions

Model Parameters	Estimated Parameters and 90% Bonferroni Confidence Intervals (C.I.)	
	Estimate	90% C.I.
Intercept, β_0	-.61	-.686 to -.545*
Stratum, β_1	.738	.689 to .789*
Letter, β_2	.227	.130 to .324*
Stamp, β_3	.090	-.006 to .186
Reminder, β_4	.291	.194 to .387*
Letter/Stamp, β_5	.036	-.101 to .173
Letter/Reminder, β_6	-.054	-.192 to .083
Reminder/Stamp, β_7	-.043	-.179 to .093
Let/Reminder/Stmp, β_8	-.003	-.197 to .191

A C.I. marked by an * indicates the difference was statistically significant at $\alpha = .10$.

The results of both modelings show that significant improvements were realized from the prenotice letter and reminder post card, but not from the stamped return envelope for the national and within stratum models. These results correspond to those presented by the multiple comparisons above. None of the interaction terms were statistically significant, indicating the effect of the components are basically additive in nature.

4. DISCUSSION AND CONCLUSIONS

The prenotice letter, stamp and reminder postcard individually improved response rates by 6.4, 2.5 and 8.0 percentage points, respectively. The increase of 2.5 was not statistically significant. The effects of the elements were also found to be mostly additive, and did not interact with one another. In comparison to the control group, the

combination of letter-stamp improved response 9.8 percentage points, the stamp and reminder, 9.5 percent, and the letter and reminder, 12.7 percent. All three elements together improved response by 14.3 percent. Each use of the letter and reminder added significantly to response, but the stamp only added significantly when used with a prenotice and no reminder. The most important conclusion from this experiment was that both the prenotice letter and reminder postcard are important to achieving a high response and that neither eliminates the effect of the other.

Although the individual effect (2.5 percent overall) of the stamped return envelope is slightly smaller than needed for significance, it is of similar magnitude to what has been found significant in past research (Armstrong and Luske 1987; Dillman 1978, 1991). In light of the preponderance of past research showing its effectiveness, this technique should probably not be completely dismissed as being ineffective. It also appears that the stamped return envelope relates differently to the prenotice and reminder. When used alone with the prenotice, the effect of the stamped return is significant (3.4 percentage points), but it is clearly insignificant (1.6 percentage points) when a reminder is included in the mailout procedures. The reminder compensates for the lack of a stamped return envelope, whereas the prenotice appears to amplify its effect. It may be that a prenotice alerts people to notice and open the census form mailout package, and once opened, people are then encouraged to respond by the presence of the stamped return envelope. This differential connection to the mailings that precede and those that follow, appears not to have been examined in past research. A practical implication for the Census is that if a prenotice letter and no reminder is used, a stamped return envelope might add significantly to response, but be of less importance if a reminder postcard is used, as was done in the last census.

There are at least two significant barriers to the direct application of this research to conduct of the 2000 Census. First, it is important to recognize that these tests are being done in non-census years. In the past the Census Bureau has obtained much lower response rates in non-census years than in census years. For example, the 1986 National Content Test, obtained only a 49.2 percent response employing a replacement questionnaire, while the 1990 Census without employing a replacement questionnaire, achieved a 65 percent response rate. The usual explanation for this difference is "census climate," a succinct explanation of the combination of media attention, advertising, and cultural sense of participation that seems to build each decade during the census year.

The response rates obtained in our tests with the use of the five elements found to increase response are much higher than normally obtained in non-census years, but are close to the same, or perhaps a little lower, than those obtained during the last decennial census when none of these elements were used. We do not know whether the

existence of a "census climate" will substitute for the effects of these elements or add to the response likely to be obtained in a census year. Certainly a 30 percentage point increase will not be realized in the 2000 Census since that would suggest a response of nearly 100 percent. Therefore, considerable uncertainty remains with respect to the exact implications of the present findings for the 2000 Census.

APPENDIX

Estimation Procedures

Analytical results are derived from two separate methods, multiple comparisons among the mail response rates by treatment group, and logistic regression analysis. Each method has advantages over the other in terms of ease of interpretation and ease of statistical inference; hence a combined approach was utilized to bring forth the best of both methods for presentation.

The national mail response rate estimates for a given panel as presented in this study is computed by dividing the weighted total of the number of questionnaires returned by the weighted total number of forms mailed out less weighted postmaster returns (mostly vacant units).

Multiple comparisons of the 8 treatment mail response rates were reviewed to determine the level of increase in the mail response to each of the treatments. These comparisons involved a pairwise assessment of each of the treatments with the control panel and with each other.

The logistic regression procedures provide a quick and effective means for evaluating whether or not observed increases from each of the components, especially interactions, are the result of sampling variation or imply a true increase, and if these increases are influenced by the presence of other components. However, parameter estimates cannot be easily equated to the mail response rates. A detailed overview of the logistic regression methodology is provided in Thompson 1993.

Response rates were calculated for each of the treatment groups within stratum and at the national level (stratum 1 and stratum 2 combined). Standard errors for the national estimates were computed using the stratified jackknife variance procedure (Wolter 1985). The estimates were produced by the VPLX statistical software package. Standard errors for the within stratum estimates were computed using the formula for the simple random sampling jackknife variance procedure.

The primary analysis involved pairwise comparisons of the differences between response rates for eight treatments, both overall at the national level and for the two strata, LRA and HRA.

Because of the various hypotheses being tested, all possible pairwise comparisons (28 total) between the eight treatments are analyzed in the experiment. In the logistic

regression framework 8 or more model parameters are tested for significance. The more comparisons that are made, the greater the potential that some of these comparisons will be incorrectly declared significant. In this case, additional statistical measures are employed to control the overall error of the decision process.

The analysis has been carried out so that statements about the entire "family" of 28 pairwise comparisons or the logistic regression parameters are made while maintaining the 90 percent (a Census Bureau standard) confidence level simultaneously for all comparisons. All 90 percent confidence intervals for the pairwise comparisons were adjusted using Dunnett's C-procedure for comparing pairwise contrasts of the test panel estimates (Hochberg and Tamhane 1987). Bonferroni simultaneous inference procedures were used to evaluate the statistical significance of the logistic regression parameters.

REFERENCES

- ARMSTRONG, J.S., and LUSKE, E.J. (1987). Return postage in mail surveys: A meta analysis. *Public Opinion Quarterly*, 51 (1) 233-248.
- DILLMAN, D.A., CLARK, J., and SINCLAIR, M. (1993). Effects of questionnaire length, respondent-friendly design, and a difficult question on response rates for occupant-addressed census mail surveys. *Public Opinion Quarterly*.
- DILLMAN, D.A., SINCLAIR, M., and CLARK, J. (1992). Mail-back response rates for simplified decennial census questionnaires. *Proceeding of the Section on Survey Research Methods, American Statistical Association*, 776-783.
- DILLMAN, D.A. (1991). The design and administration of mail surveys. *Annual Review of Sociology*, 17, 225-249.
- DILLMAN, D.A. (1978). *Mail and Telephone Surveys: The Total Design Method*. New York: Wiley-Interscience.
- DUNCAN, W.J. (1979). Mail questionnaires in survey research: A review of response inducement techniques. *Journal of Management*, 5, 39-55.
- FOX, R.J., CRASK, M.R., and KIM, J. (1988). Mail survey response rate: A meta-analysis of selected techniques for inducing response. *Public Opinion Quarterly*, 52, 467-491.
- HARVEY, L. (1987). Factors affecting response rates to mailed questionnaires: A comprehensive literature review. *Journal of the Market Research Society*, 29, 3, 342-353.
- HEBERLEIN, T., and BAUMGARTNER, R. (1978). Factors affecting response rates to mailed questionnaires: A quantitative analysis of the published literature. *American Sociological Review*, 43, 447-462.
- HOCHBERG, Y., and TAMHANE, A.C. (1987). *Multiple Comparison Procedures*. New York: John Wiley and Sons.
- KANUK, L., and BERENSON, C. (1975). Mail surveys and response rates: A literature review. *Journal of Marketing Research*, 12, 440-453.

- KULKA, R.A., HOLT, N.A., CARTER, W., and DOWD, K.L. (1991). Self reports of time pressures, concerns for privacy and participation in the 1990 Mail Census. *Proceedings of the 1991 Annual Research Conference*. U.S. Bureau of the Census, 33-54.
- LINSKY, A.S. (1975). Stimulating responses to mailed questionnaires: A review. *Public Opinion Quarterly*, 39, 82-101.
- MISKURA, S.M. (1992). Estimating the Full Cycle Costs for the Simplified Questionnaire Test (SQT), 2KS Memorandum Series, Design 2000, Book I, Chapter 30, #6.
- SCOTT, C. (1961). Research in mail surveys. *Journal of Royal Statistical Society*, 143-205.
- THOMPSON, J.H. (1993). Final Results of the Mail Response Evaluation for the Implementation Test (IT), DSSD 2000 Census Memorandum Series, #E-32.
- WOLTER, Kirk (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

Consistency of Census and Vital Registration Data on Older Americans: 1970-1990

LAURA B. SHRESTHA and SAMUEL H. PRESTON¹

ABSTRACT

Major uncertainties about the quality of elderly population and death enumerations in the United States result from coverage and content errors in the censuses and the death registration system. This study evaluates the consistency of reported data between the two sources for the white and the African-American populations. The focus is on the older population (aged 60 and above), where mortality trends have the greatest impact on social programs and where data are most problematic. Using intercensal cohort analysis, age-specific inconsistencies between the sources are identified for two periods: 1970-1980 and 1980-1990. The U.S. data inconsistencies are examined in light of evidence in the literature regarding the nature of coverage and content errors in the data sources. Data for African-Americans are highly inconsistent in the 1970-1990 period, likely the result of age overstatement in censuses relative to death registration. Inconsistencies also exist for whites in the 1970-1980 intercensal period. We argue that the primary source of this error is an undercount in the 1970 census relative to both the 1980 census and the death registration. In contrast, the 1980-1990 data for whites, and particularly for white females, are highly consistent, far better than in most European countries.

KEY WORDS: Age misreporting; Coverage; Mortality; Census evaluation; Death registration; Data quality; Mortality crossover, United States.

1. INTRODUCTION

Conventional methods of estimating levels of mortality in more developed countries use data from two different sources. The numerators of death rates are normally counts of deaths derived from vital statistics. The denominators are usually derived from census counts of persons alive. The accuracy of calculated rates depends on the quality of data from both sources.

This paper reports the results from a test of data quality applied to United States data for two intercensal periods: 1970-1980 and 1980-1990. In particular, we examine the consistency of reported changes in the size of a cohort between two censuses and the recorded number of intercensal deaths for that cohort, with allowance for intercensal cohort migration. All data refer to the population in single years of age and separate tests are conducted for the black and white populations.

Our focus is on the older population (aged 60 and above), where mortality trends have the greatest impact on social programs (Preston 1993) and where data quality is most problematic. The white population of the United States appears to have lower death rates above age 80 than any other industrialized country (Vaupel 1993). If valid, this comparison would have important implications for evaluating the relative quality of medical systems. But the African-American population of the United States has even lower rates than the white population above age 80,

reflecting the well-known crossing over of the age patterns of mortality between the races somewhere between ages 75 and 85. Whether either set of mortality rates can be accepted at face value depends, of course, on the quality of the data. Data on blacks has elicited considerable skepticism (e.g., Zelnik 1969; Coale and Kisker 1990), although most observers appear to accept the validity of the crossover (Manton *et al.* 1986; McCord and Freeman 1990).

In the process of constructing new model mortality patterns for low mortality countries, Condran, Himes, and Preston (1991) report similar data quality tests for 68 intercensal periods in 18 industrialized countries. In general, consistency was very good for cohorts aged 65 at the second census (66 of 68 data sets passed the consistency check). Consistency deteriorated with age; only about half of the data sets showed consistency at age 85 and fewer than 15% did so at age 95 (Condran *et al.* 1991: Table 7). The United States was not among the countries included in these tests because it lacked published data on deaths by single year of age. We are now able to fill in this important gap because we have processed data tapes on each individual death registered in the United States from 1970 through 1988. (The single year death distribution for 1989 (full year) and 1990, January to March only, is estimated using published group data from the National Center for Health Statistics and the 1988 single-year death distribution. Details are provided in Appendix A.) These tapes are produced by the National Center for Health Statistics

¹ Laura B. Shrestha, The World Bank, Human Development Department, 1818 H Street, N.W., Washington, DC 20433, U.S.A.; Samuel H. Preston, Population Studies Center, University of Pennsylvania, 3718 Locust Walk, Philadelphia, PA 19104, U.S.A.

(NCHS) and are distributed by the Inter-University Consortium for Political and Social Research. For the years we have included, they contain approximately 50 million deaths.

2. STUDY POPULATION AND DATA

2.1 Background

Three major sources of data are utilized: (1) national-level census enumerations from the U.S. Bureau of the Census for the years 1970, 1980 and 1990; (2) annual death registration data produced by NCHS; and (3) unpublished estimates of net immigration obtained from the U.S. Bureau of the Census. While the data sources are described in more detail in Appendix A, a brief description of the data and significant adjustments is warranted.

2.2 The Census Enumerations

We utilize census tabulations which are classified by race (black/white), sex, and single years of age (open-ended at age 100). The tabulations refer to the resident population of the 50 states and the District of Columbia. Included in the enumerations are: the institutionalized population, Americans travelling abroad temporarily, and foreign citizens having their usual residence (legally or illegally) in the United States (except foreign military and diplomatic personnel). Specifically excluded are: Americans overseas for an extended period and foreign citizens temporarily visiting the U.S. The official statistics do not adjust for census undercount, *e.g.*, the failure to find and enumerate legal residents and undocumented resident immigrants.

The term "resident population" implies that both the legal population and undocumented immigrants are included in the census tabulations. While undocumented persons were residing in the U.S. at the time of the 1970 census, it appears that only a negligible number were counted. Hence, the legal resident population approximated the total resident population in the 1970 census. In the 1980 count, however, the U.S. Bureau of the Census estimates that, for the first time ever, a significant number of undocumented persons were enumerated. Estimates indicate that the count equalled 2.06 million undocumented persons. Of this number, in the age group 60 and above, 10 thousand white males were enumerated; 19 thousand white females, 3 thousand black males, and 6 thousand black females (U.S. Bureau of the Census 1988).

The official 1970 census tabulations are known to contain errors, the most conspicuous of which is the gross overstatement of the number of persons aged 100 years or more. Although the census enumerated 106,000 persons in this age group, indirect demographic estimates indicated that the correct centenarian count should have been in the range of 3,000 to 8,000 with a preferred estimate of 4,800

(Siegel 1974; Siegel and Passel 1976; U.S. Bureau of the Census 1974). We utilize unpublished U.S. Census Bureau tabulations of the 1970 census, which correct for the centenarian overcount. Use of the corrected estimates is justified by two conditions: first, without adjustment, the excess is large enough to bias results at the oldest ages. Second, it appears that the overcount was not due to systematic misreporting of age into the centenarian population. Rather, it was the result of misunderstanding of the census form wherein individuals confused the columns intended for month of birth and year of birth (Siegel and Passel 1976).

In both the 1980 and 1990 censuses, a large number of individuals enumerated chose to write in a response to the race question as opposed to selecting one of the specified all-inclusive race categories. For the total population, 6.8 million individuals, largely of Spanish-origin, were affected in 1980, whereas the number increased to 9.3 million in the 1990 census. The official census tabulations are not directly comparable with other data sources since only the census enumerations contain a residual race category. To allow comparison with other data systems, the Census Bureau modified the 1980 and the 1990 enumerations to conform to historical categories of the racial groupings. The 1990 modification at the Census Bureau also involved "correction" for an age-related problem (for details, see Word and Spencer 1991). The decision was made to use the race-modified statistics for 1980 and 1990 from the Census Bureau for this research. The choice is justified by the sheer magnitude of individuals that would be excluded by use of the unmodified data, particularly for the white population.

2.3 Death Registration Data

The U.S. death registration data represent every death registered in the 50 states and the District of Columbia, classified by race, sex and age (single years of age to 125+). To insure comparability with the census data, deaths of nonresidents of the United States (nonresident foreign nationals and U.S. nationals residing abroad) have been excluded.

Adjustment is made for neither under-registration of deaths nor for misreporting of characteristics on the death certificates. Two problems were identified that affected the utilization of our intended intercensal methodology. The intercensal period covers the interval from April 1 to March 31, whereas the death registration data refer to calendar years. And both the death registration and the U.S. censuses' data are reported by age at last birthday rather than by year of birth. We manage both problems by assigning deaths to triangles of time-age that correspond to "census years" beginning on April 1. For example, deaths reported in the one year interval between census date April 1, 1970 and April 1, 1971 to those aged 60 (last

birthday) at the time of the census can be classified into four categories: (1) deaths to those aged 60 in calendar year 1970; (2) deaths to those aged 60 in calendar year 1971; (3) deaths to those aged 61 in calendar year 1970; and (4) deaths to those aged 61 in calendar year 1971. Using data on the date of death from the NCHS tapes, we assigned deaths to triangles of time-age that corresponded to the census year beginning on April 1, 1970. In doing so, we assume that deaths within each triangle are evenly distributed. This assumption is necessitated by the lack of reliable birth data for most of the cohorts considered in the paper, data that could be used to apportion deaths more accurately among adjacent birth cohorts. For a more detailed description of the methodology, see Shrestha 1993.

2.4 Net Immigration Statistics

We utilize unpublished net immigration statistics obtained from the U.S. Bureau of the Census. While the quality of net immigration statistics in the U.S. is widely acknowledged to be suspect (Hill 1985), the estimation of population size at the older ages is quite robust to variations in estimates of intercensal migration. This robustness results both from the smaller flow of net migrants at the older ages relative to younger ages and from the greater magnitude of deaths as a source of decrement in the older age groups relative to changes as a result of net migration. For instance, the net immigration data list an inflow of 64 black males for the cohort aged 75 and above (in 1970) during the 1970 to 1980 decade. For comparison, over 141 thousand deaths were recorded for the same cohort.

Estimates of the flow of undocumented residents are not included in the constructed net immigration series, but will be considered in the interpretation of results. Their exclusion was precipitated by a number of factors. Estimates of the size and age-sex distributions of the illegal alien population vary widely due to insufficient data collection instruments in the U.S. But even the most exaggerated estimates of the number of undocumented migrants are minuscule relative to deaths at the older ages.

We have described a number of adjustments that we have made to the basic data: use of unpublished 1970 census tabulations because of a gross overcount of the centenarian population in the official statistics, use of race-modified tabulations of the 1980 and 1990 census, and exclusion of estimates of the undocumented alien population. In order to judge the effect of these adjustments on our results, we carried out numerous sensitivity analyses using uncorrected data. While only modest differences were observed between the results using official statistics and those with corrected tabulations (except at ages 100 and above), the intercensal cohort analyses using uncorrected data generally produced greater deviations in our final results, implying the overall appropriateness of these corrections.

3. SOURCES OF ERROR IN CENSUSES AND DEATH DATA

Errors in demographic data have been classified into coverage errors and content errors. Coverage refers to the completeness with which persons or events that fall within the defined universe of a particular data system are recorded. Content refers to the quality of information about the persons or events that are in fact recorded. Either type of error in any data source can create inconsistencies between intercensal change in cohort size and intercensal deaths. However, if both censuses and death registration suffer from the same net omission rate, then the sources will be consistent with one another; but under these circumstances, recorded death rates will also be accurate.

Identical patterns of age misreporting in censuses and death registration will not, in general, produce consistency between changes in cohort size between the censuses and recorded numbers of intercensal deaths. The reason is that, because death rates rise with age, the age distribution of deaths at older ages is older than the age distribution of population. For example, if 10% of both persons and deaths at true ages 75-79 are misreported into the age interval 80-84, then the proportionate impact on population counts will be greater than the proportionate impact on death counts. Such a pattern of age misreporting would distort death rates, and would also be visible in the consistency tests that we apply.

The Census Bureau has used demographic and statistical procedures to estimate the completeness of census coverage. Demographic procedures compare estimates of the true numbers of births minus estimated cohort deaths and migrations to census counts (see the summary in Robinson *et al.* 1993 and Himes and Clogg 1992). Statistical procedures match a group of individuals identified in an alternative data source (such as the Current Population Survey) to individual-level records from the Census. A third approach is to compare the Census count of older persons to the count of individuals in Medicare files.

A number of general conclusions for the old-age population were reached in the evaluation studies of the 1970 census undertaken by the U.S. Bureau of the Census (1973, 1974, 1975). First, the magnitude of net error (combination of coverage and content errors) in the old-age statistics is greater than for the younger population. Second, females exhibited higher net error rates than males, largely the result of higher levels of age misreporting. But, gross omission rates (which are only one component of net error) were higher for males. Third, levels of net error, of gross omission, and of misreporting of demographic characteristics are considerably higher for the U.S. black population than for the white. Fourth, the evidence suggests that considerable age misreporting exists in the official statistics. For example, it is interesting to note that,

for all four race-sex groups at ages 65-69 years in 1970, the estimates derived by demographic analysis suggest net census overcounts, whereas the Medicare linkage study found gross census omissions in the magnitude of 2.1% (for white females) to 12.6% (for males of the black and other races category). This comparison implies that, while these groups have gross omissions in the number of persons enumerated at ages 65-69, other larger errors (presumably, especially age overstatement among persons below age 65) are operating in the other direction to inflate the net overcount estimates at these ages. One implication is that the characteristics of a substantial part of the population reported as 65 and over in the Census relate to persons who are in fact under age 65 (U.S. Bureau of the Census 1976).

Relative to the 1970 census, the net error rates in 1980 in most of the age-race-sex groups were significantly lower. As noted by the Bureau of the Census (1988), however, results from the Post Enumeration Program (PEP) and from the 1980 Housing Unit Enumeration Duplication study affirm that a considerable proportion of the total census count, likely in excess of 1.1%, represented duplicate enumerations of individuals already in the census. Evidence implies much lower levels of duplication in earlier censuses. Thus, "regrettably, duplication receives dubious credit for part of the improvement in 1980 in net census coverage" (U.S. Bureau of the Census 1988:10).

The Census Bureau plans exhaustive evaluations of the quality of the 1990 Census, but the release of such analyses has been fragmentary to date. It does appear that the gross undercount was lower in 1980 than in 1990 (Robinson *et al.* 1993), but this may be the result of a higher degree of duplications in the 1980 census. A number of generalizations can be made regarding the pattern of net undercount in the 1990 census for the aged population. First, following its historical trend, the net error estimates for African-Americans surpass those of whites by a wide margin. The largest differential is noted for males aged 60-64. The net undercount rate for black males equals 10.3 percent, surpassing the white male estimate of 2.6 percent by 7.7 percentage points. Second, whereas undercounts are observed for all of the male aged categories, overcounts are noted in many of the female groups. Finally, as noted by Robinson *et al.* (*ibid*), the net coverage patterns are generally consistent across the last three censuses for each race-sex group.

Official death statistics produced by the National Center for Health Statistics are the basic source of annual mortality data in the United States. The figures are generally utilized without adjustment for underregistration or for misreporting of characteristics on the death certificate. It is generally assumed, however, that the death registration system is practically complete (Wilkin 1981; U.S. Bureau of the Census 1984a; National Center for Health Statistics

1968) although no national test of its comprehensiveness has been conducted since the completion of the Death Registration Area in 1933. This assumption is based on the strict legal requirements for registration as well as on the needs of survivors for proof of death in connection with burial, settling estates and collecting insurance benefits (U.S. Bureau of the Census 1984a; Wilkin 1981). Calculations by Coale and Kisker (1990), however, suggest that underregistration of deaths exists, particularly at the older ages. For the nonwhite population, for instance, registered deaths were 7% fewer than Medicare deaths for the male population aged over 80 in 1980, whereas registered female deaths were 10% fewer. These numbers, however, may be reflective of differential age reporting between the two sources, rather than of underenumeration.

The best evidence regarding the consistency of age reporting between censuses and death registration – undoubtedly the most important source of content error affecting our consistency test – matched a sample of death certificates from May to August 1960 with the 1960 census records (NCHS 1968; Hambricht 1969). Although the data were collected before the time frame considered for this project, the study's findings provide insight into what may be a continuing pattern of biases present in the census and death statistics. The authors found: (1) for whites, there was fairly high agreement between the sources even with increasing age – for nonwhites, however, there was less agreement; (2) in the event of disagreement, age discrepancies for the white population between the sources were generally within one year – for nonwhites, however, the typical difference was more than one year, particularly at ages 45 and above; and (3) for whites of all ages and nonwhites aged less than 45 years, the age reported on the death certificate was typically older than that reported on the census – for nonwhites aged 45 and above, however, age reported on the death certificate was, on average, younger than on the census.

This study was unable to ascertain which data source, if either, provides the "true" age. To this end, Rosenwaike and Logue (1983) attempted to verify age reporting on the death certificate for the population aged 85 and over in the 1968 to 1972 period. The authors selected a sample of death records from those filed for decedents of extreme age in Pennsylvania and New Jersey. They then linked the individual who died to the 1900 manuscript census of population. A total of 1429 decedents were linked of whom 960 were white and 496 were non-white.

They found that age agreement of matched census records with death certificates decreased as age increased for both racial groups. Striking differences were noted between racial groups. Agreement levels for whites were high, except at ages 100 and over. For nonwhites, however, significantly lower agreement was found. The authors further note that, within race, there was little difference by sex in agreement on age.

4. AN INTERCENSAL METHODOLOGY TO EVALUATE THE QUALITY OF OLD-AGE STATISTICS

This analysis examines the extent of inconsistency in old-age U.S. data sources using an intercensal cohort methodology. The expected size of an open-ended age cohort in the second census can be estimated from its size at the first census and the intercensal deaths occurring to that cohort, after adjustment for migration (Condran, Himes and Preston 1991). Use of an open-ended category allows observation of the ratio trend while dampening error-induced extreme values at particular ages. It is insensitive to any errors of age reporting in deaths or population that occur within the population above the age that begins the open-ended age interval.

Using census enumerations and death and migration statistics for an intercensal period, intercensal cohort analysis allows us to estimate the expected size of each open-ended age cohort in the subsequent census. The previously mentioned statistics, classified by single years of age, by sex, and by two races (white, black), were utilized to calculate the following equation for the expected population at the time of the second census:

$$\hat{N}_x(2) = N_{x-10}(1) - D_{x-10}(1) + M_{x-10}(1) \quad (1)$$

where

$\hat{N}_x(2)$ = the predicted population aged x and above at the second census, taken 10 years after the first.

$N_{x-10}(1)$ = the enumerated population aged $x - 10$ and above at time 1, the first census.

$D_{x-10}(1)$ = the intercensal deaths which had occurred to the cohort aged $x - 10$ and above (at the first census).

$M_{x-10}(1)$ = intercensal net legal immigration into the cohort aged $x - 10$ and above (at the first census).

Similarly, the expected population at a given age (as opposed to at age x and above) can be calculated in an analogous manner. In either circumstance, the ratio of the observed population, enumerated in the subsequent census, to the expected population, can then be calculated (after simplifying the notation and assuming net migration to be zero) as:

$$R_x = \frac{N_x(2)}{N_{x-10}(1) - D} \quad (2)$$

The change in the size of the cohort as measured at two successive censuses can be produced only by death or migration. A ratio of 1.00 would indicate complete consistency among the data sources. (Note that a ratio of 1.00,

while highlighting consistency, does not assure accuracy. On an individual level, for instance, if a person's age was consistently overstated by n years, the method would fail to capture the misreporting.) In fact, however, the reported count will also be affected by: (1) coverage errors in either or both censuses; (2) under- (or over-) enumeration in the death registration data and/or the immigration statistics; and (3) misreporting of characteristics (age, race, *etc.*) in any or all of the data sources (Ewbank 1981; Shryock and Siegel 1976; Condran *et al.* 1991). The ratio of observed to expected population is a useful diagnostic tool if patterns of deviation from 1.00 can be interpreted in terms of these underlying data errors. It is not a highly precise tool because different forms of error can produce the same pattern of ratios. Nevertheless, it can help discriminate among competing alternatives.

5. HOW PATTERNS OF ERROR WILL AFFECT OBSERVED/EXPECTED RATIOS

Effects of certain types of error are visible directly in the formula for the ratio itself (and have been confirmed by simulations that we have performed). To simplify the exposition, define R_x in equation (2) as the ratio of observed to expected population for age x at the second census. The following major possibilities for coverage error, and their implications for the age-pattern of ratios, can be distinguished:

- 1) If $N_{x-10}(1)$ and D are equally complete and $N_x(2)$ has a relative completeness level of $C(2)$, then the age pattern of ratios will be constant with age and its level will be $C(2)$.
- 2) If $N_x(2)$ and D are equally complete and $N_{x-10}(1)$ has a relative completeness level of $C(1)$, then the age pattern of ratios will be:
 - a) Above 1.00 and rising with age if $C(1) < 1.00$
 - b) Below 1.00 and falling with age if $C(1) > 1.00$.

The reason why an age trend in R_x results from this pattern of error is that a particular proportionate error in $N_{x-10}(1)$ creates increasingly larger proportionate errors in the denominator as the two offsetting terms (one positive and one negative) in the denominator grow more equal in absolute value. This equalization occurs because a higher fraction of each cohort dies during the intercensal period as age advances.

- 3) If $N_{x-10}(1)$ and $N_x(2)$ are equally complete and D has a relative completeness level of $C(D)$, then the age pattern of ratios will be:
 - a) Above 1.00 and rising with age if $C(D) > 1.00$ (*i.e.*, if completeness of death registration exceeds the completeness of enumeration in both censuses).
 - b) Below 1.00 and falling with age if $C(D) < 1.00$.

Once again, an age trend is introduced because an equal proportionate error in D will create larger proportionate errors in the denominator as its two components become more equal in absolute value.

Some of the effects of age misreporting patterns can also be understood by examining the components of this formula. Shrestha (1993) and Condran *et al.* (1991) introduce various errors into simulated errorless data sets typical of the current demographic conditions of the United States and the Netherlands respectively. They show that a pattern of net overstatement of age that is confined to the two censuses will produce a pattern of ratios that hovers around 1.00 until advanced ages, whereupon it falls to very low values. The reason why the ratio declines below 1.00 is, once again, that an error in one component of the denominator (in this case, inflation of $N_{x-10}(1)$ by age overstatement) introduces disproportionate effects in the denominator. Even though the rapid tapering off in the age distribution can result in $N_x(2)$ being more inflated than $N_{x-10}(1)$, eventually the inflation of the denominator exceeds that of the numerator and the ratios fall. (For an illustration, see Figure 1 of Condran *et al.* 1991).

Age overstatement that is confined to deaths will create a pattern of ratios that is above 1.00 and rises with age; the denominator is too low (its negative component is too large) and the proportionate deficit grows with age.

Introducing the *same* pattern of age overstatement into deaths and population figures also creates ratios that eventually rise with age. This important result is robust to the extent of error introduced (Condran *et al.* 1991). It reflects the fact that age distributions taper off more and more rapidly as age advances, so that the *same* percentage of persons who overstate their true age will introduce larger *percentage* errors in the reported age distributions at the very advanced ages. That is, $N_x(2)$ has a larger inflation factor than $N_{x-10}(1)$. In this case, some inflation in $N_{x-10}(1)$ is offset in its effects on the denominator by an inflation in D .

6. RESULTS

Intercensal cohort analysis was carried out for the four sex-race groups in the United States in the 1970-1980 and 1980-1990 periods. Figures 1 and 2 present the calculated ratios of the observed to expected population at selected ages by race, sex, and intercensal period.

In all race-period combinations, the age pattern of ratios is virtually the same for females and males. In all cases, the degree of inconsistency increases with age, although any systematic and significant departure from 1.00 is postponed until age 95 and beyond for whites in 1980-1990. There is clearly a discontinuity in many of these series at age 100, reflecting the idiosyncrasies of age reporting and Census Bureau adjustment procedures among centenarians.

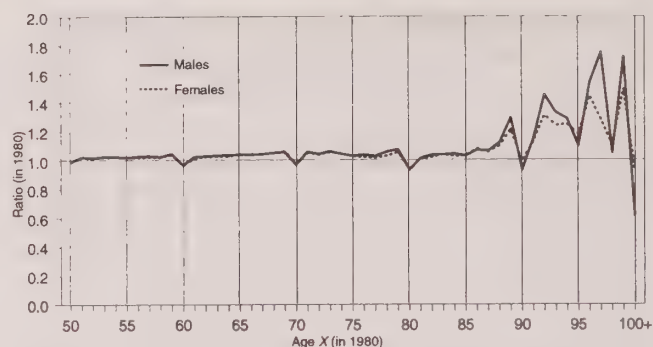


Figure 1A. Intercensal Ratios of Observed to Expected Population: Whites, 1970-1980.

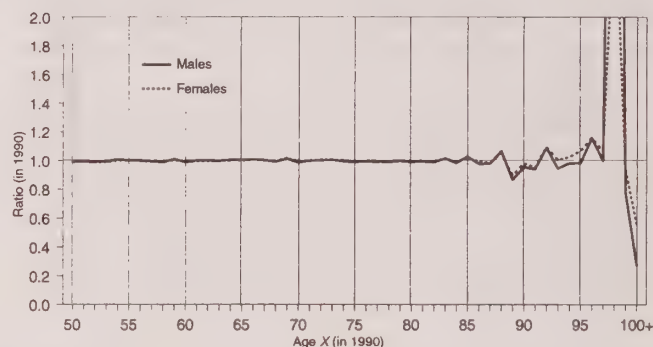


Figure 1B. Intercensal Ratios of Observed to Expected Population: Whites, 1980-1990.

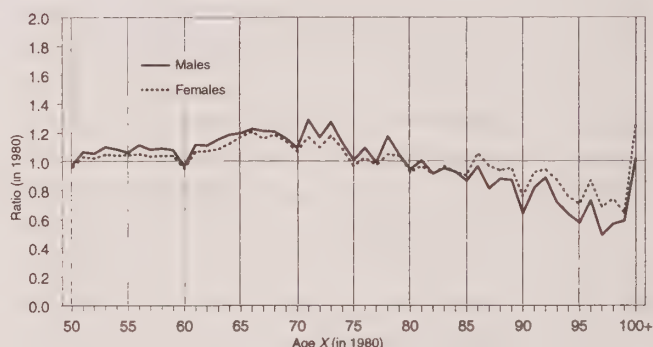


Figure 2A. Intercensal Ratios of Observed to Expected Population: Blacks, 1970-1980.

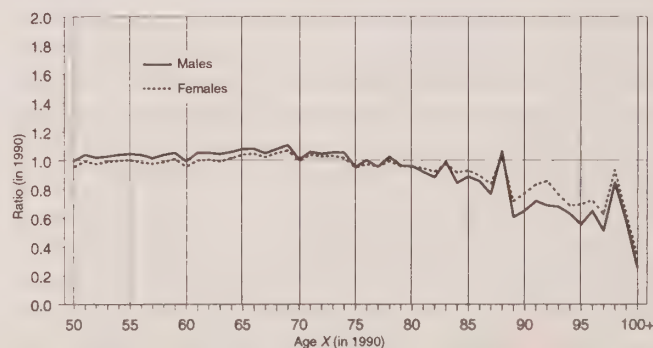


Figure 2B. Intercensal Ratios of Observed to Expected Population: Blacks, 1980-1990.

6.1 Results for Whites

6.1.1 Intercensal Period: 1970-1980

As shown in Figure 1A, the white pattern in 1970-1980 is generally above unity and rising with age (up to age 100). This pattern is consistent with several forms of data error, the two most plausible patterns of which are:

- 1) Undercount in the 1970 census, relative to both the 1980 census and the death registration.
- 2) Roughly equal probabilities of age overstatement in deaths and in both censuses.

We believe that the former explanation is more likely to be correct. If the pattern of ratios resulted from similar tendencies for age misstatement in deaths and censuses, one would expect that pattern to continue into the 1980-1990 decade, particularly since the 1980 census is involved in both comparisons. And one would not expect cultural predispositions to misstate age to disappear suddenly. But the 1980-1990 pattern of ratios for whites (Figure 1B) shows remarkable consistency, far better than that in most European countries and equivalent to the pattern of ratios found in Sweden and the Netherlands, countries with highly efficient population registers (Condran *et al.* 1991). The consistency during 1980-1990 is also much greater than that in other English-speaking countries: England and Wales, Canada, Australia and New Zealand.

A second reason for accepting the first explanation is that the Census Bureau has concluded that the 1980 census is more complete than the 1970 census (U.S. Bureau of the Census 1988; Robinson *et al.* 1993). This conclusion is partially based on demographic analysis and hence is not entirely independent of the kind of evidence that we are reviewing. However, their demographic analysis is weighted heavily towards ages that are younger than those considered here. Furthermore, the conclusion that census coverage improved is also supported by their post-enumeration program in which individuals in the census are matched against other data systems.

6.1.2 Intercensal Period: 1980-1990

As noted earlier, the 1980-1990 pattern of ratios for whites (and particularly for white females) is highly consistent, far better than in most European countries. Our investigation seemingly lends support to Vaupel's (1993) contention that the white population of the United States may have lower death rates above age 80 than any other industrialized country. But caution is in order. While our methods clearly highlight the consistency between the censuses and the death registration in 1980-1990, consistency is not equivalent to accuracy. Condran *et al.* (1991) demonstrate one situation in which a pattern of age misreporting can result in a ratio series at exactly 1.00 at all ages. Furthermore, the intercensal methodology fails to

capture deliberate misreporting of age by individuals that is consistent over time. As noted by Horiuchi (1993), an initial overstatement of age – *e.g.*, to allow entrance into school or the labor force at a younger age, to avoid being drafted near the upper limit of drafting age, or to receive Social Security, Medicare, or pension payments earlier – may be followed by consistent, intentional overstatement of age. The possibility of such overstatement of age cannot be discounted although we are unable to measure it directly.

6.2 Results for Blacks

In contrast, the pattern of ratios for African-Americans is far more regular over time (see Figure 2A and 2B). The ratios begin falling around age 70 for both sexes in both periods and continue falling through higher ages (until age 100 in 1970-1980). Before age 70, ratios are typically well above unity in 1970-1980, and slightly above 1.00 for African-American males during 1980-1990.

The fact that ratios are generally higher for African-Americans at a particular age in 1970-1980 than in 1980-1990 is consistent with a relative undercount in the 1970 census. As we noted earlier, such an undercount is also likely to have occurred among whites. The undercount, however, is insufficient to explain the persistent pattern of falling ratios above age 70 in both periods. The declining ratio series for African-Americans is consistent with two principal explanations:

- 1) Deaths are underregistered for the African-American population relative to completeness of census coverage.
- 2) Age overstatement is greater in censuses than in death registration.

Coale and Kisker (1990) lean toward the former explanation. They note that populations reconstructed from deaths using variable-*r* procedures (Preston and Coale 1982) are too small relative to census counts in 1980 above age 65, suggesting relative underregistration of deaths. They also note that fewer African-American deaths are recorded at advanced ages in vital registration than in Medicare records.

However, both observations are also consistent with ages being overstated in censuses (and Medicare) relative to death registration. That such a pattern exists is strongly supported by a direct match of death certificates in 1960 to records for the same individuals in the 1960 census of population (NCHS 1968; Hambricht 1969). For either males or females, the total number of deaths above age 50 when deaths are classified according to census age are within 1% of the total number of deaths when classified according to death certificate age. However, at ages 65 +, "census age" deaths are 15.4% greater than "death certificate age" deaths for females and 7.1% greater for males. At age 75 +, the disparities are 23.3% and 17.8%, respectively, and at age 85 +, 39.2% and 17.6%.

These large discrepancies in age reporting between censuses and deaths are capable of accounting for the declining pattern of ratios above age 70 that is demonstrated in Figure 2. Elo and Preston (1994) calculate the R_x values for African-Americans between 1950-1960 and 1960-1970, periods that bound the 1960 census-death certificate match. They show that, if ages at death are "corrected" to make them consistent with the age reporting in the censuses, the pattern of declining ratios is eliminated.

Reasons why African-American ages are overstated in censuses relative to deaths are not obvious. The pattern does not appear until the 1940 census, the first census after Social Security legislation was passed. At that census, a large surplus of African-American persons aged 65-69 and 70-74 appears, and a deficit of persons aged 50-64 (Elo and Preston 1994). As noted by Wolfenden (1954:56), "the disturbances were so marked in the data for Negroes that special preliminary redistributions of those populations (and deaths) between 55 and 69 were made in the preparation of the [U.S.] life tables." This surplus also appears, although in increasingly attenuated form, in more recent censuses (as shown in Figure 2). Whatever its source, we believe that the principal explanation of the large inconsistencies between censuses and death registration for the African-American population is a pattern of age overstatement in censuses relative to death registration. Such a pattern implies that recorded death rates above age 65 for African-Americans are likely to be seriously underestimated. A cross-over between black and white death rates may indeed occur at advanced ages, but basing such a conclusion on U.S. census and vital registration data is treacherous. These data are simply too inconsistent with one another to allow death rates at advanced ages to be estimated with any confidence.

7. CONCLUSION

Major uncertainties about the quality of elderly population and death enumerations in the United States result from coverage and content errors in the censuses and the death registration system. This study evaluates the consistency of reported data between the two sources for the white and the African-American populations. The focus is on the older population (aged 60 and above), where mortality trends have the greatest impact on social programs and where data are most problematic. Using intercensal cohort analysis, age-specific inconsistencies between the sources are identified for two periods, 1970-1980 and 1980-1990.

In order to evaluate what combinations of coverage completeness and age misreporting patterns would produce the empirical results, a series of simulations were carried out. The U.S. data inconsistencies are examined in light of both the simulation results and evidence in the literature

regarding the nature of coverage and content errors in the data sources.

Data for whites in the 1980-1990 intercensal period were found to be remarkably consistent. Data quality up to age 95 approaches that of Sweden and the Netherlands, countries which maintain highly efficient population registers. Less consistency was observed for whites during the 1970-1980 decade. The most likely explanation for this pattern of inconsistencies is the relative net undercount in the 1970 census combined with more complete death statistics. Consequently, mortality estimates at older ages that combine numerators from the death registration with denominators from the 1970 census are likely to overstate mortality.

A different pattern is observed in the African-American data. Above age 70, the enumerated population falls increasingly below the expected population in both 1980 and 1990. It appears that the major reason for this pattern is that ages are overstated in censuses relative to death registration. Such a pattern implies that recorded death rates at older ages for African-Americans are likely to be seriously underestimated. A mortality crossover between black and white death rates may occur at advanced ages, but basing such a conclusion on census and vital registration data is hazardous.

ACKNOWLEDGEMENTS

This research was carried out at the University of Pennsylvania, Population Studies Center, and was supported by a grant from the National Institute of Aging, AG10168, and from the Boettner Institute of Financial Gerontology. For helpful comments on the project and/or paper, we are indebted to Irma Elo, Douglas Ewbank, Shiro Horiuchi, and J. Gregory Robinson. We are especially grateful to J. Gregory Robinson and the U.S. Bureau of the Census for supplying unpublished census and international migration tabulations.

APPENDIX A

Source: Shrestha (1993)

Three major sources of data were utilized in this research: (1) census enumerations for 1970, 1980, and 1990; (2) official death registration data; and (3) net immigration statistics. Sources of the data and adjustments made will be described.

1.A The 1970 Census

Official tabulations of the 1970 population by basic demographic characteristics are presented in *Series B - U.S. Summary of the 1970 Census* (U.S. Bureau of the Census 1972). The official enumerations are known to

contain a number of major inaccuracies which could bias our investigation of the enumerated old-age population in the United States. The first is a conspicuous overcount of the centenarian population. Whereas 106,000 persons were enumerated in the open-ended category, indirect demographic analysis estimates the correct count to be in the range of 3,000 to 8,000 (Siegel 1974; Siegel and Passel 1976). The overcount appears to have been the result of misunderstanding of the census form rather than systematic age misreporting into the centenarian population. The second problem is the result of misclassification of the population by race in the complete-count tabulations, affecting 21,000 individuals aged 65 and above. And finally, the official count omitted over 23,000 individuals (of all ages) whose records were discovered after the initial tabulations were published.

Because of the inherent errors in the official tabulations, we utilize unpublished adjusted tabulations obtained from the U.S. Bureau of the Census. The modified statistics include corrections for the three previously mentioned problems. The data are presented by race (white, black), sex, and age (single years of age 0-94 and grouped data 95-99, 100+). To distribute the grouped data from age group 95-99 to single years of age, we used the sex- and race-specific average age distribution from the 1960 and 1980 censuses for whites, and from the 1950 and 1980 censuses for blacks (data by single years of age is not available in this age range for blacks in the 1960 census).

1.B The 1980 Census

Originally published census tabulations for 1980 were presented in *Series B – U.S. Summary of the 1980 Census* (U.S. Bureau of the Census 1983). In the 1980 census, however, a large number (about 6.8 million) of persons enumerated chose to write-in a response to the race question as opposed to selecting one of the specified all-inclusive race categories. Since only the 1980 census contained a residual race category, the official enumeration was not directly comparable with other data sources (vital registration, earlier censuses, etc.). The Census Bureau produced a modified file which conforms to the historical categories of the racial groupings (U.S. Bureau of the Census 1984b). The modification procedure involved macro-level reassignment of race based on detailed cross-tabulation of race and Hispanic-origin from the sample and complete-count census data. The specifics of the Census Bureau modification follow.

For the 219.8 million individuals who chose one of the 14 specified categories, no adjustment was made. Two categories of individuals, totalling 6.7 million, with write-in responses were identified: persons of Hispanic-origin (5.8 million) and persons not of Hispanic-origin (0.9 million). Separate adjustment procedures for the two groups were developed.

Those of Hispanic-origin were distributed only to the white or black categories (and not to American Indian or Asian/Pacific Islander categories). All persons of Mexican origin were reassigned as white. Persons of Puerto Rican, Cuban, and other Spanish origin were assigned to both white and black modified race groups on the basis of the distribution of the same Hispanic-origin individuals who originally specified either a white or black race on the census returns. The calculations were carried out within age-sex-county cells.

Those not of Hispanic-origin were reassigned to all three modified race groups (white, black, other) on the basis of state-specific proportions which are applied to all age-sex-county cells within the state. The proportions are based on sample data from the 1980 census. For a more detailed discussion of the modifications, see U.S. Bureau of the Census 1984b.

The modified tabulations are presented by race, sex, and single years of age (0-99; 100+). We utilize the race-modified statistics in this research, justified by the sheer magnitude of persons transferred from the residual race category to the white or black categories.

1.C The 1990 Census

Published tabulations of the 1990 Census continue to be released by the U.S. Bureau of the Census. The published statistics, however, contain a number of problems that make comparability with earlier censuses and other sources of data difficult. Three problems are apparent: racial classification of 9.3 million individuals in a residual non-specified racial category, inconsistencies in the reporting of age, and a change in allocation procedures for the 1990 census in assigning age to persons with missing data on the characteristic.

A modified 1990 census file, referred to as the MARS (Modified Age and Race Statistics) was produced at the Census Bureau to adjust for the first two problems (Word and Spencer 1991). Modification of the 1990 census was conducted at the micro-level. Hot-deck imputation procedures were utilized to assign a specific race to persons who reported themselves in the "other, not specified" racial category. The method is executed on the individual records of the 100% edited detail file from the 1990 census (Robinson, Word and Spencer 1991).

We again utilize the modified statistics, which are tabulated by race, sex, and single years of age, in this research. The decision to use the modified statistics in both 1980 and 1990 was not clear-cut. See Shrestha 1993 for a more detailed discussion.

2. The Death Registration System

National-level annual death statistics from the National Center for Health Statistics (NCHS) are utilized in this research. The data for 1970 through 1988 are extracted

from NCHS data tapes obtained from ICPSR (NCHS 1970-1988). The data are provided by race (black, white), sex, and single years of age (0-124; 125+). Since the data tapes for calendar year 1989 and for the first three months of 1990 had yet to be released, we developed a procedure to estimate the distribution. Final mortality statistics for 1989 by race and sex were released in published form by NCHS (1992). The grouped age data was distributed to single years of age based on the 1988 death distribution within the grouped age category. Distribution to month of death was based on monthly vital statistics reports (NCHS 1989). Estimates of the death distribution in 1990 are based on monthly advance reports of mortality from NCHS (1990). The preliminary numbers were distributed to single years of age again using the 1988 distribution within the grouped age category.

As noted in the text, we adjusted the available data to correct for two problems. First, the intercensal period covers the interval from April 1 to March 31, whereas the death registration data refer to calendar years. Second, both sets of data are reported by age at last birthday rather than by year of birth. Because the census is on April 1, the latter is preferred because it identifies the birth cohort for use in cohort analysis. To adjust for these two problems, we assume that the three dimensional surface of the number of deaths in age and time is level over the interval. We do not adjust for underregistration nor for misreporting of characteristics in the death statistics.

3. Net Immigration Statistics

We utilize unpublished net immigration statistics obtained from the U.S. Bureau of the Census. The tabulations are categorized in the form of "components of change" for each of the two decades.

Age-, race-, and sex-specific net immigration was calculated on a cohort basis by use of the following equation:

$$\begin{aligned} \text{Net immigration} = & \text{Legal Alien Immigration} + \text{Refugees} \\ & + \text{Parolees} + \text{Net Civilian Citizens Immigration} \\ & + \text{Net Puerto Rican Immigration} + \text{Net Foreign} \\ & \text{Students Immigration} + \text{Net Movement of U.S.} \\ & \text{Armed Forces Overseas} - \text{Legal Emigration.} \end{aligned}$$

Given the lack of sufficient detail in the raw data provided by the U.S. Census Bureau, a number of adjustments were required. First, the data had been provided with an early terminal age group (age 75 and above at the beginning of the decade). To distribute to five-year age groups (75-79, ..., 95-99, 100+), we assumed that the age-, race-, and sex-specific net immigration rate for ages 75+ remained constant in the open-ended interval beginning at age 75. This admittedly crude estimate is adequate because of small numbers of net immigrants in this age group. Second, to convert the five-year data into single years of age, we used Sprague multipliers or osculatory interpolation (Sprague 1880-81).

REFERENCES

- COALE, A.J., and KISKER, E.E. (1990). Defects in data on old-age mortality in the United States: new procedures for calculating mortality schedules and life tables at the highest ages. *Asian and Pacific Population Forum*, 4(1), 1-31.
- CONDRAN, G.A., HIMES, C.L., and PRESTON, S.H. (1991). Old-age mortality patterns in low-mortality countries: an evaluation of population and death data at advanced ages, 1950 to present. *Population Bulletin of the United Nations*, 30, 23-60.
- ELO, I.T., and PRESTON, S.H. (1994). New estimates of old-age mortality among African-Americans, 1930-1990. *Demography*, 31(3), 427-58.
- EWBANK, D.C. (1981). *Age Misreporting and Age-Selective Underenumeration: Sources, Patterns, and Consequences for Demographic Analysis*. National Academy of Sciences, Committee on Population and Demography. Report No. 4. Washington, DC: National Academy Press.
- HAMBRIGHT, T.Z. (1969). Comparison of information on death certificates and matching 1960 census records: age, marital status, nativity, and country of origin. *Demography*, 6(4), 413-24.
- HILL, K. (1985). Illegal aliens: an assessment. In: Panel on Immigration Statistics. *Immigration Statistics, A Story of Neglect*. Washington, DC: National Academy Press.
- HIMES, C.L., and CLOGG, C.C. (1992). An overview of demographic analysis as a method for evaluating census coverage in the United States. *Population Index*, 58(4), 587-607.
- HORIUCHI, S. (1993). Personal communication dated April 20, 1993.
- MANTON, K.G., STALLARD, E., and VAUPEL, J.W. (1986). Alternative models for the heterogeneity of mortality risks among the aged. *Journal of the American Statistical Association*, 81, 635-644.
- MC CORD, C., and FREEMAN, H.P. (1990). Excess mortality in Harlem. *New England Journal of Medicine*, 322, 172-177.
- NATIONAL CENTER FOR HEALTH STATISTICS (1968). *Comparability of age on the death certificate and matching census record: United States - May - August 1960*; Vital and Health Statistics: Data Evaluation and Methods Research. By Thea Zelman Hambright. Series 2, No. 29.
- NATIONAL CENTER FOR HEALTH STATISTICS (1970-1988). *Mortality Detail Files* (data and codebooks). Data were made available by the Inter-University Consortium for Political and Social Research, University of Michigan.
- NATIONAL CENTER FOR HEALTH STATISTICS (1989). Births, marriages, divorces, and deaths for January-December 1989. *Monthly vital statistics report*. Vol. 38, Nos. 1-12. Hyattsville, Maryland: Public Health Service.
- NATIONAL CENTER FOR HEALTH STATISTICS (1990). Births, marriages, divorces, and deaths for January-March 1990. *Monthly vital statistics report*. Vol. 39, Nos. 1-3. Hyattsville, Maryland: Public Health Service.

- NATIONAL CENTER FOR HEALTH STATISTICS (1992). Advance report of final mortality statistics, 1989. *Monthly vital statistics report*. Vol. 40, No. 8, supp. 2. Hyattsville, Maryland: Public Health Service.
- PRESTON, S.H. (1993). Demographic change in the United States, 1970-2050. *Demography and Retirement: The 21st Century*, (Sylvester Scheiber, Ed.). New York: Praeger Press.
- PRESTON, S.H., and COALE, A.J. (1982). Age structure, growth, attrition, and accession: a new synthesis. *Population Index*. 48(2), 217-59.
- ROBINSON, J.G., AHMED, B., DAS GUPTA, P., and WOODROW, K.A. (1993). Estimation of population coverage in the 1990 United States census based on demographic analysis. *Journal of the American Statistical Association*, 88, 1061-1079.
- ROBINSON, J.G., WORD, D.L., and SPENCER, G. (1991). Uncertainty for models to translate 1990 census concepts into historical racial classifications. 1990 Decennial Census, Preliminary Research and Evaluation Memorandum (PREM) No. 81, Demographic Analysis Evaluation Project D8.
- ROSENWAIKE, I., and LOGUE, B. 1983. Accuracy of death certificate ages for the extreme aged. *Demography*, 20(4), 569-585.
- SHRESTHA, L.B. (1993). Age Misreporting and its Effects on Old-Age Population and Death Registration Estimates: United States, 1970-1990. Unpublished doctoral dissertation, University of Pennsylvania, Population Studies Center.
- SHRYOCK, H.S., and SIEGEL, J.S. (1976). *The Methods and Material of Demography*. Orlando, Florida: Academic Press, Inc. (Harcourt Brace Jovanovich, Publishers): Studies in Population Series.
- SIEGEL, J.S. (1974). Estimates of coverage of the population by sex, race, and age in the 1970 census. *Demography*, 11(1), 1-23.
- SIEGEL, J.S., and PASSEL, J.S. (1976). New estimates of the number of centenarians in the United States. *Journal of the American Statistical Association*. 71, 559-566.
- SPRAGUE, T.B. (1880-81). Explanation of a new formula for interpolation. *Journal of the Institute of Actuaries*. 22, 270.
- UNITED STATES BUREAU OF THE CENSUS (1972). *General Population Characteristics*. 1970 Census of Population and Housing. Final Report PC(1)-B1. U.S. Summary. Washington, DC: U.S. Government Printing Office.
- UNITED STATES BUREAU OF THE CENSUS (1973). *The Medicare Record Check: an Evaluation of the Coverage of Persons 65 Years of Age and Over in the 1970 Census*. 1970 Census of Population: Evaluation and Research Program, PHC(E)-7. Washington, DC: U.S. Government Printing Office.
- UNITED STATES BUREAU OF THE CENSUS (1974). *Estimates of Coverage of Population by Sex, Race, and Age: Demographic Analysis*. 1970 Census of Population: Evaluation and Research Program, PHC(E)-4. By Jacob S. Siegel. Washington, DC: U.S. Government Printing Office.
- UNITED STATES BUREAU OF THE CENSUS (1975). *Accuracy of Data for Selected Population Characteristics as Measured by the 1970 CPS-Census Match*. 1970 Census of Population: Evaluation and Research Program, PHC(E)-11. Washington, DC: U.S. Government Printing Office.
- UNITED STATES BUREAU OF THE CENSUS (1976). *Demographic Aspects of Aging and the Older Population in the United States*. Current Population Reports. Series P-23, No. 59. By J.S. Siegel. Washington, DC: U.S. Government Printing Office.
- UNITED STATES BUREAU OF THE CENSUS (1983). *General Population Characteristics*. 1980 Census of Population and Housing. Final Report PC80-1-B1. United States Summary. Washington, DC: U.S. Government Printing Office.
- UNITED STATES BUREAU OF THE CENSUS (1984a). *Demographic and Socioeconomic Aspects of Aging in the United States*. Current Population Reports. Series P-23, No. 138. By J.S. Siegel and M. Davidson. Washington, DC: US Government Printing Office.
- UNITED STATES BUREAU OF THE CENSUS (1984b). Census of Population: 1980. Race detail file. 100% count. Table IV: modified counts (OMB-consistent) by age, race, and sex. Unpublished tabulations.
- UNITED STATES BUREAU OF THE CENSUS (1988). *The Coverage of Population in the 1980 Census*. 1980 Census of Population and Housing: Evaluation and Research Reports, PHC80-E4. By R.E. Fay, J.S. Passel and J.G. Robinson. Washington, DC: U.S. Government Printing Office.
- VAUPEL, J.W. (1993). Verbal presentation at Research Workshop on Oldest Old Mortality, Duke University, Durham, North Carolina. March, 1993.
- WILKIN, J.C. (1981). Recent trends in the mortality of the aged. *Transactions of the Society of Actuaries*. Vol. XXXIII, 11-62.
- WOLFENDEN, H.H. (1954). *Population Statistics and their Compilation*. Revised Edition. The University of Chicago Press: published for the Society of Actuaries.
- WORD, D.L., and SPENCER, G. (1991). Age, sex, race, and Hispanic origin information from the 1990 census: a comparison of census results with results where age and race have been modified. 1990 CHS-L-74. Draft dated August 1991.
- ZELNIK, M. (1969). Age patterns of mortality of American Negroes: 1900-02 to 1959-61. *Journal of the American Statistical Association*. 64, 433-451.

An Assessment of the Use of Hand-Held Computers During Demographic Surveys in Developing Countries

D. FORSTER and R.W. SNOW¹

ABSTRACT

Although large scale surveys conducted in developing countries can provide an invaluable snapshot of the health situation in a community, results produced rarely reflect the current reality as they are often released several months or years after data collection. The time lag can be partially attributed to delays in entering, coding and cleaning data after it is collected in the field. Recent advances in computer technology have provided a means of directly recording data onto a hand-held computer. Errors are reduced because in-built checks triggered as the questionnaire is administered reject illogical or inconsistent entries. This paper reports the use of one such computer-assisted interviewing tool in the collection of demographic data in Kenya. Although initial costs of establishing computer-assisted interviewing are high, the benefits are clear: errors that can creep into data collected by experienced field staff can be reduced to negligible levels. In situations where speed is essential, a large number of staff are involved, or a pre-coded questionnaire is used to collect data routinely over a long period, computer-assisted interviewing could prove a means of saving costs in the long term, as well as producing a dramatic improvement in data quality in the immediate term.

KEY WORDS: Hand-held computers; Demographic surveys; Psion.

1. INTRODUCTION

Large scale surveys involving tens of thousands of respondents, such as national censuses, demographic or health surveys, are routinely conducted in developing countries. Their intention is to provide rapid, up-to-date information on population and health issues for evaluation and planning purposes. Their wide scope necessitates numerous personnel comprising trainers, interviewers, supervisors, data entry staff and data managers. Examples of such questionnaire-based surveys include the World Fertility Survey (WFS 1986) and national Demographic and Health Surveys (DHS Kenya 1989). Published dates for the commencement of the WFS surveys in 12 African countries and the dates the first country reports were produced (Table 1) illustrate the time required before data was available for planners to act upon (WFS 1986). On average it took 45.6 months before the final report was released. Survey logistics in developing countries undoubtedly contribute to delays in provision of completed data; so do the mechanics of data processing. The recent Demographic and Health Survey conducted in Kenya required five data entry clerks, two data entry supervisors and a control clerk to process 8,343 household interviews; data collection began in February 1989 and the first draft of the final report was ready for circulation seven months later (DHS Kenya 1989).

Table 1

Summary of Chronology of 12 African WFS Surveys
(Source: WFS 1986)

Country	Number of Interviews	Date Survey Started	Date of First Report	Number of Months from Survey Start Till Report Date
Benin	4,018	12/1981	06/1984	30
Cameroon	8,219	01/1978	04/1983	63
Ghana	6,125	02/1979	06/1983	52
Ivory Coast	6,270	08/1980	12/1984	52
Kenya	8,100	08/1977	06/1980	34
Lesotho	3,603	08/1977	12/1981	52
Mauritania	3,500	01/1981	06/1984	41
Morocco	5,800	04/1980	05/1984	49
Nigeria	9,727	10/1981	09/1984	35
Senegal	3,985	05/1978	07/1981	38
Sudan (North)	3,115	12/1978	04/1982	40
Tunisia	4,123	05/1978	06/1983	61

¹ D. Forster, Department of Tropical Medicine, University of Oxford, John Radcliffe Hospital, Headington OX3 9DU, England; R.W. Snow, CRC - Research Unit, Kenyan Medical Research Institute, P.O. Box 230, Kilifi, Kenya.

Surveys of this size involve multiple levels of checking and coding of data collected in the field providing another source of delay. As speed underpins rapid health evaluation (Anker 1991; Vlassoff and Tanner 1992), reducing the time at this check and code stage is a major advantage to the survey process. Advances in computer hardware have led to the development of microcomputers suitable for use in field situations. Together with improved software designed for questionnaire specification and administration, computer-assisted interviewing is now a viable option. National statistics offices in industrialised countries have evaluated the use of this technique, and some now use them on a regular basis (Nicholls and Groves 1986; Lyberg 1985; Denteneer *et al.* 1987; Bench *et al.* 1994). The advantages of these systems are that it reduces recording errors by simplifying skip modules and refusing inappropriate, illogical or inconsistent entries. Furthermore, large numbers of interviews can be stored and simply downloaded to a central computer at the end of every interview session, circumventing the need for data entry clerks.

There is surprising reluctance to adopt this technology in developing countries despite its apparent advantages. There are several possible reasons for this. Firstly, the initial costs may seem daunting and the application deemed inappropriate in countries with scarce resources. Secondly, there have been few attempts to validate their use under field conditions providing little quantifiable evidence of their limitations or advantages over traditional data collection techniques (Reitmaier 1985; Ferry and Cantrelle 1988; Forster *et al.* 1991). This paper presents the results of a comparative study of two methods of field data collection and processing conducted during a demographic survey on the Kenyan Coast.

2. THE ADULT MORTALITY SURVEY

The study was carried out as part of ongoing demographic and epidemiological studies of 60,000 people living on the Kenyan coast. The study population and survey methods employed to monitor demographic events has been described elsewhere (Snow *et al.* 1994). In brief, following an initial census of the population all vital events are monitored by means of 6-weekly house-to-house visits and bi-annual re-censuses of the entire population. During a re-enumeration of the population in November 1993, a survey was undertaken to estimate adult mortality using indirect demographic methods (Timaues 1991). All women aged between 25 and 44 years were interviewed using the structured questionnaire as shown in Figure 1. The format used precoded closed questions, with logical skips and a consistency check.

Twenty-four field staff, all secondary school leavers, were involved in the survey. All were familiar with survey and census procedures, having had previous formal training

in field survey techniques and between 1 and 5 years of field experience. Two days was spent on additional training on the administration of the adult mortality questionnaire. During the survey field staff were divided into two teams, each supervised by a senior fieldworker. Questionnaires completed at the end of each day were checked by field supervisors then passed to the computer staff for data entry. This was done using a screen design reflecting the structure of the paper questionnaire in FoxPro (version 2.0). The same data was independently entered by two data entry clerks and the two completed files compared to identify entry errors, which were subsequently corrected. The completed file was then subjected to logical, range and consistency checks; these included for example, the identification of missing data, incorrect coding (*i.e.*, not using "Y" or "N"), dates inconsistent with the ages of the women and the date of the survey (questions 5 and 6 in Figure 1) and checks that the sums of questions 7, 9, 11 and 13 are consistent with question 15 as shown in Figure 1.

3. COMPUTER DATA COLLECTION TEST

3.1 Computer Hardware and Software

An earlier version of questionnaire-based software was developed for the Psion Organiser II (Forster *et al.* 1991). This model had a limited screen size, 16 characters by 2 lines, but had a fully operational keyboard. The Psion Series 3, used during the present study, offers new possibilities: the screen is much larger, with 40 characters by 8 lines, and integrated graphical capabilities. The machine remains small (165mm by 85mm by 22mm), and weighs 265g including 2 AA sized batteries. The storage devices can store up to 1 megabyte. The keyboard is a 58 key, QWERTY layout. Communications between the Psion Series 3 and a PC entails a simple copy operation between the two storage media.

The software was developed using Psion's in-built programming language, OPL. The paper questionnaire is represented in a structured format in a text file, according to a prescribed format. The questionnaire definition includes a mixture of questions and commands such as skips and range checks. The internal range checks included those developed for the inconsistency checks for the data entered using FoxPro described above. Data entered on the Psion is stored in a separate file, one line for each interview.

To specify a question correctly it must include a question number, the question text and the answer type, which can be a list option, a character input or a number. The definition should also indicate what position in the line the corresponding data entry should be stored and how long the entry is. Numeric answers can also include a prespecified number of decimal points. A range of acceptable inputs

Figure 1. The Adult Mortality Questionnaire

Questionnaire on the survival of relatives			
(For all women aged 25-44 years)			
Names _____			
Date _____	ID _____-_____-_____		
I would like to ask you some questions about your natural parents and about your brothers and sisters who have the same mother as you.			
1.	Is your mother alive? (1 = yes, 2 = no)		<input type="text"/>
2.	Is your father alive? (1 = yes, 2 = no)		<input type="text"/>
<i>INTERVIEWER: If both parents alive (Q1 and Q2 = 1), go to Q7.</i>			
3.	Have you ever given birth? (1 = yes, 2 = no)		<input type="text"/>
<i>INTERVIEWER: If she has never given birth (Q3 = 2), go to Q6.</i>			
4.	Was (MENTION ALL PARENTS NOT ALIVE NOW) alive at the time that you gave birth to your first child?		
		Yes	No
	Woman's mother	1	2
	Woman's father	1	2
5.	In what year was your first child born?		<input type="text"/>
6.	In what year (MENTION ALL PARENTS NOT ALIVE NOW) die?		
	Woman's mother		<input type="text"/>
	Woman's father		<input type="text"/>
7.	How many living sisters, born to your mother, do you have? (ALIVE NOW)		<input type="text"/>
<i>INTERVIEWER: If no living sisters (Q7 = 0), go to Q9.</i>			
8.	How many of these living sisters are less than 15 years old?		<input type="text"/>
9.	How many of your sisters, born to your mother, have died?		<input type="text"/>
<i>INTERVIEWER: If no dead sisters (Q9 = 0), go to Q11.</i>			
10.	How many of these dead sisters died before age 15 years?		<input type="text"/>
11.	How many living brothers, born to your mother, do you have? (ALIVE NOW)		<input type="text"/>
<i>INTERVIEWER: If no living brothers (Q11 = 0), go to Q13.</i>			
12.	How many of these living brothers are less than 15 years old?		<input type="text"/>
13.	How many of your brothers born to your mother, have died?		<input type="text"/>
<i>INTERVIEWER: If no dead brothers (Q13 = 0), go to Q15.</i>			
14.	How many of these dead brothers died before age 15 years?		<input type="text"/>
<i>INTERVIEWER: Sum Q7, 9, 11 and 13:</i>			
	Q7	=	<input type="text"/>
	Q9	=	<input type="text"/>
	Q11	=	<input type="text"/>
	Q13	=	<input type="text"/>
15.	I want to make sure that I have this right. Apart from you, your mother had children altogether? Is that correct?		<input type="text"/>
<i>INTERVIEWER: In the case of any inconsistency, probe and correct Q7 to Q14 if necessary.</i>			
<i>INTERVIEWER: Please thank the woman for her co-operation.</i>			
Fieldworker code			<input type="text"/>

is an optional specification for numeric or character answers and will include a minimum, a maximum or both. List options can be used to specify codes and their values.

Command actions can be embedded in question texts, so that they are evaluated at the time of questionnaire administration. For example, the final cross-check question in Figure 1 requires an addition. The syntax allows this instruction to be included within the main body of the question text. Other commands can contain instructions on skipping a question, conducting a cross-check between answers or for moving to a different question.

Thus the software uses a flexible way of defining a questionnaire, which is generally applicable. It incorporates manipulation of entered information, integrating arithmetic functions into questions or command lines. The next step forward would be to design an interface for questionnaire specification which removes the burden of constructing a syntactically correct questionnaire definition. The software is available from the authors.

3.2 Test Design

An additional day's training on the use of the Psion was provided for the two team leaders. This involved an explanation of the hardware and software, as well as practice sessions in the field. Both supervisors had had no previous computing experience. Both conducted interviews using either the Psion or paper questionnaires on alternate days, and these formed the basis of the comparison between the methods.

Errors made by all the 22 fieldworkers using the paper questionnaire were counted and tabulated to estimate the background error rate using this method of data collection. Times taken to check the forms once they had been brought back from the field, and for the data to be entered, verified and corrected following range and consistency checks were recorded throughout the survey. Similar timing assessments were made for the Psion data collection procedures.

4. TEST RESULTS

4.1 Time

The average length of interviews conducted on paper was 5.1 minutes, and for those on computer 5.0 minutes, demonstrating no difference between the two methods (Table 2; note only 215/234 interviews were timed). The length of interviews varied considerably from 1 to 18 minutes and increased time related not simply to the number of skips made on each interview, but whether the respondent gave clear, non-contradictory answers.

Team supervisors required between 2-3 hours per day to check each teams questionnaires. The average time taken to enter 500 records (approximate number of interviews completed per week) was 3 hours 40 minutes per data

Table 2
Comparison Between Paper Questionnaire and Psion
Series 3 Data Collection Methods

	Length of Interview in Minutes				
	Minimum Overall	Maximum Overall	Average Overall	Average Leader A	Average Leader B
Paper	1	16	5.1 (215)*	5.2 (128)	4.9 (87)
Computer	1	18	5.0 (363)	5.5 (190)	4.5 (173)

* Number of timed interviews stated in brackets.

entry clerk; double entry required 7 hours 20 minutes (Table 3). Verification required an additional 2 hours 23 minutes for the same number of questionnaires. The completed files were edited twice to reflect corrected errors and then verified; this took on average two hours 30 minutes for 500 records.

Table 3
Times for Data Processing
500 Questionnaires

Activity	Average time
Data checking	4 hours 8 minutes
First data entry	3 hours 40 minutes
Second data entry	3 hours 40 minutes
Verification	2 hours 23 minutes
Editing	2 hours 33 minutes
Total time	18 hours 24 minutes

4.2 Errors

The errors made were divided into two periods, to assess the effect of familiarity over time. Excluding the two team leaders, the remaining 22 fieldworkers made 1,704 errors on 1,427 questionnaires in the first period, and 1,049 errors on 1,158 questionnaires in the second period. Thus the average error rate per questionnaire in the first two weeks was 1.19 and in the third and fourth weeks was 0.90. In addition, over the entire period 37 questionnaires (1.2% of all interviews) had to be sent back to the field to be redone, as the errors found were not reconcilable in the office. These questionnaires had between 1 and 6 errors to be corrected, with a total of 61 errors. The highest number of errors were made on question 5 (17 errors) and question 6b (15 errors). Error rates per question are shown in Table 4. Fourteen out of the 22 fieldworkers redid at least one questionnaire. One fieldworker was required to redo 8 questionnaires.

Table 4
Type of Errors Made by 22 Fieldworkers
Using Paper Questionnaires
(for question specification see Figure 1)

	Period 1 (first fortnight)	Period 2 (second fortnight)
Identification	163	48
Question 1	6	1
Question 2	8	2
Question 3	125	92
Question 4a	201	138
Question 4b	151	93
Question 5	105	61
Question 6a	94	57
Question 6b	65	41
Question 7	14	0
Question 8	109	63
Question 9	51	10
Question 10	178	134
Question 11	13	1
Question 12	108	71
Question 13	53	3
Question 14	204	149
Question 15	19	76
Fieldworker code	37	9
Total errors	1,704	1,049
Total questionnaires	1,427	1,158

Errors were detected either manually by final checking by one of the investigators (Forster) or through the range and consistency checks performed in FoxPro on the entered data. Field team leader A made 8 errors on 144 questionnaires (0.06 errors per questionnaire), and team leader B made 18 errors on 90 questionnaires (0.20 errors per questionnaire). Most of these errors occurred in question 10 (4 errors) and question 15 (5 errors). The only errors found from the computer data were errors of respondent identification. There were 7 of these, 2 by leader A and 5 by leader B, giving errors of 0.01 and 0.03 per questionnaire respectively. Such errors could have been circumvented by pre-loading the Psion with a call list of respondents to interview.

4.3 Cost

The differential costs of a survey of this size using Psion-based and paper-based methods are given in Table 5. The Psion prices quoted are the recommended retail prices. Intense competition between retailers means that purchase prices could be up to 20% lower than those quoted here.

Prices of hardware products are also decreasing. Current prices indicate that the one off cost of a Psion-based system can be recouped after 12-15 similar paper-based surveys of approximately 7,000 respondents.

Table 5
A Comparative Study of Computer-based and Paper-based
Survey Methods (UK £ Sterling)

	Equipment required	Cost
Computed-based survey	20 Psion Series 3	2,539.00
	20 1 MB storage devices	2,039.00
	1 serial communications link	59.45
	80 rechargeable batteries	146.20
	1 battery recharger	15.95
	Total cost	4,799.59
Paper-based survey	14 reams of paper for 7,000 interviews	42.00
	Duplicating costs for 7,000 questionnaires (double-sided)	70.00
	20 pens, erasers and correcting fluid	27.40
	20 clipboards	100.00
	2 data entry clerks (two weeks)	70.00
	2 supervisors* (one month plus overtime)	85.00
	Total cost	394.40

* Necessary for the manual checking of forms as they come in from the field each day.

5. DISCUSSION

The lowest error rates using a traditional paper questionnaire by senior field workers with five years of data collection experience was on average 0.11 errors per questionnaire with 17 fields. This was reduced to negligible levels using the questionnaire software developed for a Psion Series 3 hand-held computer. This technique eliminated most of the errors made by fieldworkers in the routing of the questionnaire (Table 4) by using pre-defined skip modules, thus reducing the error rate by at least 90%. With the additional implementation of a call list in the software, the rate of respondent identification errors would be even lower.

The field supervisors were keen to use the computer, mastering the unfamiliar QWERTY keyboard, and learnt operating procedures quickly enough to take to the field without supervision after two days. Although no formal investigations were undertaken to gauge and quantify interviewees' reactions to the Psion there were surprisingly few comments about the computer and no interview refusals.

Data processing involved two data entry clerks using two IBM machines full time for 92 hours to complete the data entry process for the entire survey. A data manager was on hand to offer assistance where necessary and design

the data entry format. The setting of the present study was such that both data entry clerks were familiar with data entry procedures and the available hardware and software. In situations where this is not the case, closer supervision and involvement by a data manager would be necessary, thus incurring an additional cost. A Psion data collection system would require much less of a data manager's time to down-load each day's data, thereby reducing this component of staff costs. Never-the-less, the initial cost of the Psion Series 3 may be prohibitively expensive when compared to the costs of paper and duplication of questionnaires if it was not envisaged that they form part of future data collection activities.

QUESTOR (Ferry and Cantrelle 1988) offers a suitable software environment for computer-assisted interviewing. However, the hardware required is a portable PC, several times the costs of a hand-held Psion. Our experience demonstrates that it will be worth pursuing the development of an appropriate package using this compact PC compatible technology as a more practical alternative in the field, being easier to handle, more robust and with reduced power consumption.

There is a trade off between error rates, time and cost of a survey. The use of computer-assisted interviewing software can reduce both the error rates and the length of time for data preparation considerably. Such a collection system should reduce the unacceptable delays in first presentation of data experienced during surveys such as the World Fertility Survey (Table 1). The context of the present comparative study differs from many large scale demographic surveys where recruited fieldstaff are unfamiliar with questionnaire procedures. We feel that the results presented here therefore represent a minimum improvement that could be expected in data quality. The initial cost of setting up such a survey mechanism may be daunting, but will be proportionally less for repeated surveys, or in institutions conducting a variety of different surveys over time.

ACKNOWLEDGEMENTS

The authors wish to thank all the field staff involved during the survey especially the two supervisors, Rodgers Chisengwa and Lewis Mitsanze, the data entry clerks Robert Mutai and Monica Omondi. We also acknowledge the support of Dr. Ian Timaeus, who designed the adult mortality questionnaire and Dr. Chris Nevill, who conducted the re-enumeration. We also wish to thank Dr. Kevin Marsh for his support for the computer studies at Kilifi, the Wellcome Trust, UK for financial support; the donation of a Psion Series 3 by Psion PLC, UK; and the Director of KEMRI for permission to publish this work. Dr. Bob Snow is supported as part of The Wellcome Trust Senior Fellowship programme in Basic Biomedical Science.

REFERENCES

- ANKER, M. (1991). Epidemiological and statistical methods for rapid health assessment. *World Health Statistics Quarterly*, 44, 94-97.
- BENCH, J., CLARK, C., DUFOUR, J., and KAUSHAL, R. (1994). Computer-assisted interviewing for the labour force survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*.
- DHS, KENYA (1989). Report on Kenyan Demographic and Health Survey. Institute for Resource Development/Macro Systems Inc., Columbia, Maryland, USA.
- DENTENEER, D., BETHLEHEM, J.G., HUNDEPOOL, A.J., and KELLER, W.J. (1987). The BLAISE System for Computer-Assisted Processing, Automation in Survey Processing. Netherlands Bureau of Statistics, CBS Select No. 4, 67-76.
- FERRY, B., and CANTRELLE, P. (1988). The use of microcomputers for collection of demographic data in the field. In *African Population Conference*, Dakar, Senegal. Liege, Belgium: International Union for the Scientific Study of Population (IUSSP), 15-30.
- FORSTER, D., BEHRENS, R.H., CAMPBELL, H., and BYASS, P. (1991). Evaluation of a computerized field data collection system for health surveys. *Bulletin of the World Health Organisation*, 69, 107-111.
- FORSTER, D., and SNOW, R.W. (1992). Using microcomputers for rapid data collection in developing countries. *Health Policy and Planning*, 7, 667-71.
- LYBERG, L. (1985). Plans for computer-assisted data collection at Statistics Sweden. *Bulletin of the International Statistical Institute*, 45th session, Invited Papers, Volume LI, Book 3, Section 18.2.
- NICHOLLS, W.L., and GROVES, R.M. (1986). The status of computer-assisted telephone interviewing: Part I - Introduction and impact on cost and timeliness of survey data. *Journal of Official Statistics*, 2, 93-115.
- REITMAIER, P., DUPRET, A., and CUTTING, W.A.M. (1987). Better health data with a portable microcomputer at the periphery: An anthropometric survey in Cape Verde. *Bulletin of the World Health Organisation*, 65, 651-657.
- SNOW, R.W., MUNG'ALA, V.O., FORSTER, D., and MARSH, K. (1994). The role of the district hospital in child survival on the Kenyan Coast. *African Journal of Health Sciences*, 1, 71-75.
- TIMAEUS, I.M. (1991). Measurement of adult mortality in less developed countries: a comparative review. *Population Index*, 57, 552-568.
- VLASSOFF, C., and TANNER, M. (1992). The relevance of rapid assessment to health research and interventions. *Health Policy and Planning*, 7, 1-9.
- WORLD FERTILITY SURVEY (1986). Final report. International Statistical Institute, Netherlands.

Statistical Process Control of Sampling Frames

A.W. SPISAK¹

ABSTRACT

Statistical process control can be used as a quality tool to assure the accuracy of sampling frames that are constructed periodically. Sampling frame sizes are plotted in a control chart to detect special causes of variation. Procedures to identify the appropriate time series (ARIMA) model for serially correlated observations are described. Applications of time series analysis to the construction of control charts are discussed. Data from the United States Department of Labor's Unemployment Insurance Benefits Quality Control Program is used to illustrate the technique.

KEY WORDS: Autocorrelation; ARIMA models; Control charts; Quality assurance.

1. INTRODUCTION

The integrity of the sampling frame is of paramount importance in survey research. Frame imperfections include missing elements (incomplete frame), element clusters (more than one element in a single listing), blank or foreign elements, and duplicate listings. These imperfections can cause several difficulties by contributing to nonsampling error, reducing the number of sample cases from subclasses of the population, and requiring the use of complex weights to estimate population characteristics. Techniques to minimize frame problems or reduce their impact on the survey are discussed in detail in most textbooks on statistical surveys.

This article focuses on the statistical process control of sampling frames which are constructed periodically (daily, weekly, or monthly, for example) and which consist of elements that are generated by a continuous process. Because of the variation inherent to any dynamic process, the sizes of the sampling frames will vary. How do we know that the changes in the sizes of the sampling frames reflect the random variation of the process and not errors in the construction of the frames? Statistical process control allows survey managers to distinguish between the variation inherent in the process (common causes) and variation which signals a possible problem with frame construction (special causes).

2. PROCESS VARIATION AND STATISTICAL PROCESS CONTROL

Over the last several years managers in the manufacturing, service, and public sectors of the economy increasingly have adopted the quality philosophies developed by W. Edwards Deming, J.M. Juran, Philip B. Crosby, Kaoru Ishikawa, and others. Quality management comprises an

array of tools and techniques, including the use of control charts to determine if a process is in statistical control. According to Deming (1982), statistical control is achieved by eliminating special causes of variation, leaving only the random variation of a stable process. The behavior of a process that is in statistical control is predictable.

The distinction between common and special causes of variation is a key principle of statistical process control. Deming (1982) credits Dr. Walter A. Shewhart, who developed many of the principles of statistical process control in the 1920s and 1930s, with originating the concept of special or assignable causes. Special causes are usually attributable to one part of the process, such as a worker, machine, or office. They will reoccur unless they are identified and eliminated. Special causes are signaled by data points that fall outside of the control limits, by consecutive points that fall above or below the process average, or by runs of increasing or decreasing points.

Common causes of variation are inherent to the process; they are present at all times and effect the entire process. Common causes are reduced or eliminated through management actions that change the process.

3. STATISTICAL PROCESS CONTROL APPLICATION TO THE CONSTRUCTION OF SAMPLING FRAMES FOR PERIODIC SURVEYS

3.1 United States Unemployment Insurance Benefits Quality Control

The use of statistical process control as a quality management tool for sampling frames is illustrated by an example from the United States Department of Labor's Unemployment Insurance Benefits Quality Control program. Since 1987, the 50 states, the District of Columbia, and

¹ A.W. Spisak, Mathematical Statistician, Unemployment Insurance Service, U.S. Department of Labor, Washington, DC 20210, U.S.A.

Puerto Rico have conducted the Benefits Quality Control program in cooperation with the United States Department of Labor. The goal of the program is to reduce the overpayment and underpayment of Unemployment Insurance benefits by identifying the causes of payment errors and initiating measures to improve the benefit payment process.

When an individual files a claim for Unemployment Insurance benefits, Unemployment Insurance staff determine whether the claimant has met all of the eligibility requirements – for example, the claimant earned sufficient wages in his or her previous employment to qualify for benefits; the claimant is involuntarily unemployed; and the claimant is able and available to work and is actively seeking employment. If all of the eligibility requirements are satisfied, the state Unemployment Insurance agency issues a benefits check for the week of unemployment claimed.

3.2 Benefits Quality Control Sampling Procedures and Sources of Error

Each state selects weekly random samples of Unemployment Insurance payments that are examined to determine if the correct amount was paid to the claimant. If the amount paid was incorrect, the investigator identifies the types and causes of the errors so that program managers can initiate corrective measures. The sampling frames are constructed each week from the universe of Unemployment Insurance payments that were issued between 12:00 am Sunday and 11:59 pm the following Saturday. A computer program edits the state's database to insure that only payments that meet the program's operational definition of the target population are included in the frame. For example, payments for some temporary or small Unemployment Insurance programs are excluded from the frame.

The volume of Unemployment Insurance checks issued each week (and therefore the size of the sampling frames) varies in response to the number of individuals who claim and receive benefits during that week. However, there are several sources of potential errors which can affect the integrity of the frame. Some of the most serious of these errors are:

- The payments made from some of the local Unemployment Insurance offices might not be picked up for inclusion in the state's central database, due to telecommunication or ADP problems.
- If the state builds a separate file for each day's transactions, the transactions for one or more days might be erroneously omitted from the final cumulative file.
- Incorrect coding of transactions could result in either foreign elements being included in the frame or the editing out of transactions that should be included.

4. DATA ANALYSIS AND MODEL DEVELOPMENT

Figure 1 is a time series plot of sampling frame sizes for a 52 week period. Each week's sampling frame consists of the previous week's Unemployment Insurance benefit recipients who continue to receive benefits, minus the previous week's Unemployment Insurance recipients who have returned to work, exhausted their benefits, or failed to file a claim, plus newly eligible claimants and eligible claimants who did not file a claim or were not compensated for a claim the previous week.



Figure 1. Number of UI payments per week.

Control charts for individual observations assume that the data are independent and identically distributed (i.i.d.). However, if the data are serially correlated, the estimates of the process variance (and therefore the control limits) could be seriously in error. So, before control charts for the Unemployment Insurance sampling frame data can be constructed, we have to determine if the observations are serially correlated.

The plot of the time series in Figure 1 provides *visual* evidence that the observations are not independent. The sampling frame data display distinct trends of increasing values during the first 13-week quarter, decreasing values over the next two quarters, and increasing values during the final 13-week quarter. The serial correlation suggested by the plot of the data in Figure 1 can be tested using methods developed to analyze time series. Although a detailed discussion of the analysis of time series data is beyond the scope of this article, the concepts of stationarity and autocorrelation will be examined, in order to explain the procedures used to identify the appropriate model. Readers who are unfamiliar with the basic principles of time series analysis should consult one of the many texts on the subject, in particular Box and Jenkins (1976).

4.1 Stationarity

We can think of the individual observations that constitute a time series as a collection of jointly distributed random variables – $p(z_1, \dots, z_n)$ – where p is a probability density function and z_1, \dots, z_n are random variables. If the joint distribution of the random variables does not vary with respect to time, that is, $p(z_t, \dots, z_{t+n}) = p(z_{t+m}, \dots, z_{t+n+m})$, the process is said to be *strictly* stationary. In practice strict stationarity is difficult to establish. In this application, the time series is assumed to be *weakly* stationary. This is also referred to as second-order stationarity, because the first and second moments of the process are invariant with respect to time – $E(z_t) = E(z_{t+m})$, $VAR(z_t) = VAR(z_{t+m})$, and $COV(z_t, z_{t+k}) = COV(z_{t+m}, z_{t+k+m})$.

Throughout the rest of this article, the terms *stationary* or *stationarity* refer to a process that satisfies the conditions of weak stationarity.

4.2 Autocorrelation

In a stationary time series the covariance between any two observations depends only on the number of time periods (lags) that separate them – $COV(z_t, z_{t+k}) = COV(z_{t+m}, z_{t+k+m})$. The correlation of z_t and z_{t+k} equals $COV(z_t, z_{t+k}) / VAR(z_t)$ and is denoted ρ_k , where k is the number of periods between observations. For example, ρ_1 is the correlation of observations in the time series separated by one period and equals $COV(z_t, z_{t+1}) / VAR(z_t)$. A correlation for period k is referred to as an autocorrelation, because it is the correlation for observations which constitute a time series. The autocorrelations for the various lags can be displayed in a graph called a correlogram, which is useful in identifying the appropriate model for a time series.

4.3 Time Series Model Identification

Figure 2 is the correlogram for the 52 week time series of the number of Unemployment Insurance payments in the sample frames. The autocorrelations decrease or “die out” very slowly, which is characteristic of a nonstationary process. (Again, the reader is referred to Box (1976) and other texts on time series for a complete discussion of model identification.)

One method to transform a nonstationary series to a stationary series is *differencing*. The symbol B is the backshift operator, which when applied to z_t shifts the subscript back one period. Thus, the first difference of z_t is $(1 - B)z_t = z_t - z_{t-1}$.

Figure 3 is the time series of the differences $z_t - z_{t-1}$ of the Unemployment Insurance sampling frame data. This series appears stationary around a mean of zero. (The estimated sample mean of the differences is 150.8, with a standard error of 2064.0. The test statistic $t = (150.8 - 0) / 2064$ equals .07, and the hypothesis that $\mu = 0$ cannot be

rejected). First differences might not be sufficient to achieve stationarity for other time series, and transformations such as second differences – $(1 - B)^2 z_t = (z_t - z_{t-1}) - (z_{t-1} - z_{t-2})$, seasonal differences, or logarithmic or other variance stabilizing procedures may be required.

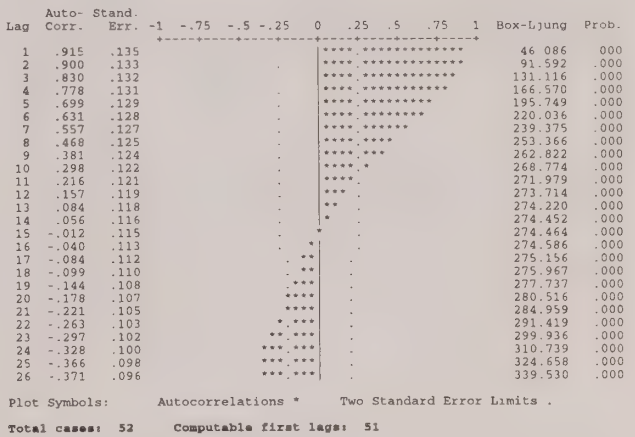


Figure 2. Autocorrelations for UI weeks paid time series.

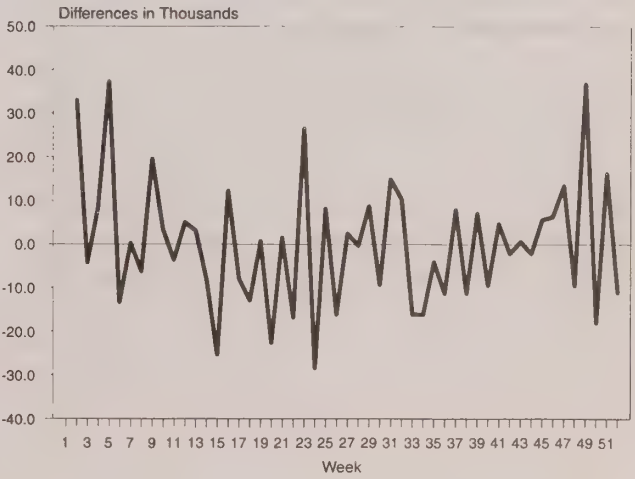


Figure 3. First differences of UI payments.

The autocorrelations of the first differences of the time series, which are displayed in Figure 4, are consistent with a stationary process. The autocorrelations decrease rapidly, while the partial autocorrelations (not displayed) die off after lag 1. This suggests that the data can be modelled with a first-order integrated autoregressive process, $ARI(1,1)$. The AR term indicates that a single autoregressive parameter will be estimated, and the integration term (I) shows that the original time series has been transformed using first differences.

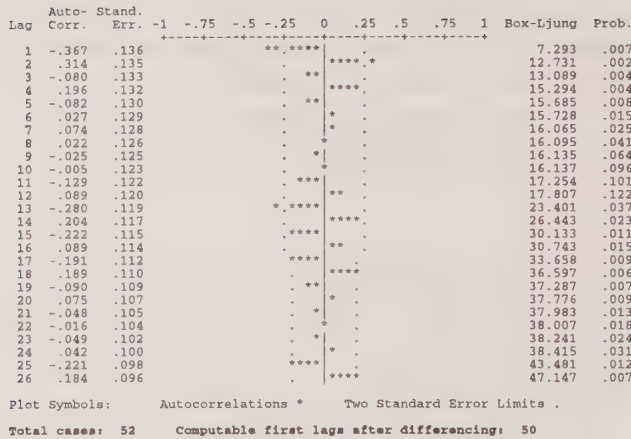


Figure 4. Autocorrelations for first differences of UI weeks paid.

4.4 Model Estimation

The model was estimated using the ARIMA procedure of the SPSS Trends software (release 4.0), which is based on the work of Box and Jenkins.

The tentative model is:

$$z_t = (1 + \phi_1)z_{t-1} - \phi_1 z_{t-2} + e_t, \text{ or}$$

$$z_t - z_{t-1} = \phi_1(z_{t-1} - z_{t-2}) + e_t,$$

where ϕ_1 is the first-order autoregressive parameter, and e is the error term, which is assumed to be normally distributed with a mean of 0 and variance σ_e^2 . The estimated autoregressive parameter, ϕ_1' is $-.4045$, and the estimated residual variance, σ_e^2 , is 184,275,853 (with 50 degrees of freedom). The negative sign on the AR parameter is consistent with the alternating signs of the autocorrelations in Figure 4. The model does not include a constant term, because the estimated process mean was not significantly different than zero.

4.5 Model Diagnostics

The adequacy of the estimated model for the observed data can be assessed by examining the model residuals. If the model adequately fits the data, the residuals (e_t) should be "white noise", that is, uncorrelated. Figure 5 displays the autocorrelations of the model residuals. Although the autocorrelation at lag 13 in Figure 5 is significant, the Box-Ljung Q statistic through lag 13 is not significant. (The Q statistic tests the significance of autocorrelations for lags 1 through k . For a detailed discussion, see Box and Pierce (1970)). In addition, none of the partial autocorrelations (not displayed) are significant. These results indicate that the residuals are not serially correlated.

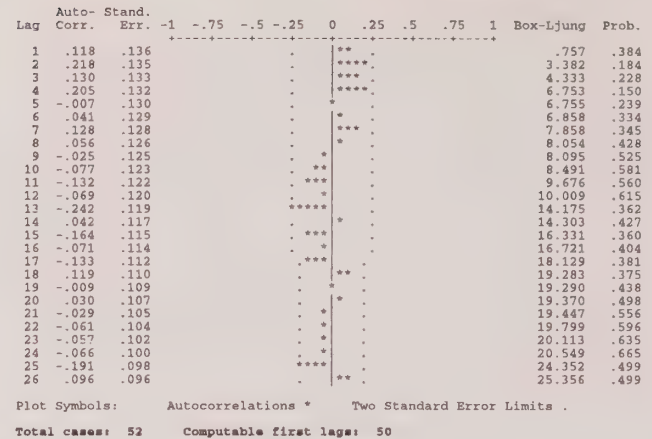


Figure 5. Autocorrelations for time series model residuals.

To test the assumption that the model residuals are normally distributed, $N(0, \sigma_e^2)$, a Kolmogorov-Smirnov ($K-S$) goodness of fit test was conducted. For the estimated variance of 184,275,853, the $K-S$ test statistic equals .591 ($p = .876$), and the hypothesis that the differences are normally distributed cannot be rejected.

For a stationary AR (1) process, the absolute value of the autoregressive parameter must be less than one. To test the hypothesis that $|\phi_1| \geq 1$ for the model, we compute: $t = (|\phi_1'| - 1)/SE(\phi_1')$, where $|\phi_1'|$ is the absolute value of the estimated autoregressive parameter, and $SE(\phi_1')$ is the standard error of ϕ_1' . The model statistics result in $t = (.4045 - 1)/.1295$ or $t = -4.6$. The chance of observing an absolute value of ϕ_1' as small as .4045 if the true absolute value of $\phi_1 \geq 1$ is very small ($< .00001$). The hypothesis that $|\phi_1| \geq 1$ is rejected, and we can conclude that the series of first differences is stationary.

5. USE OF THE ARIMA MODEL IN A CONTROL CHART

5.1 Control Charts for Individual Observations

The control limits for a chart of individual observations are set at $\bar{x} \pm 3\sigma'$, where \bar{x} is the average of observation values and σ' is the estimated standard deviation of the process. Ryan (1989) discusses alternative procedures to estimate the process standard deviation either by computing the average of the moving ranges (the mean of the absolute differences of successive observations) or using the standard deviation (s) of the sample observations, $\sigma' = s/c$, where c is an adjustment constant which depends on the sample size.

When data are serially correlated, the use of either the sample standard deviation or the average moving range can result in poor estimates of σ . The control limits constructed from these estimates can produce seriously

misleading results by either generating false signals that the process is out of control or failing to detect special causes of process variation. The moving range can underestimate σ , because the differences of successive values will tend to be small if the successive observations are highly correlated. The underestimation of σ will result in control limits that are too narrow and an increase in the number of signals of special causes. Ryan notes that using the sample standard deviation to estimate the process standard deviation will result in a better estimate of σ than the average moving range when the data are correlated, provided the sample consists of at least 50 observations. However, the sample standard deviation is an unbiased estimator of σ only when the observations are independent.

Vasilopoulos and Stamboulis (1978) analyzed the effect of serially correlated data on the control limits of \bar{x} and s (standard deviation) charts and developed equations for factors that can be used to adjust the control limits for data generated by an autoregressive process. Alternatively, a time series model can be identified for the correlated data, and a control chart can be constructed using the model residuals to monitor the process. This approach is described by Berthouex, Hunter, and Pallesen (1978) for subgroups of measurements of environmental data collected at water treatment plants. Alwan and Roberts (1988) use the residuals of exponentially weighted moving average (EWMA) models for both stationary and nonstationary time series. Montgomery and Mastrangelo (1991) use the residuals of an autoregressive model in an EWMA chart and contend that EWMA charts can be used to approximate many autocorrelated models, particularly if the observations are positively correlated and the mean does not drift too quickly. The reader is also referred to Maragah and Woodall (1992) and Woodall and Faltin (1993) for additional discussion of the effects of autocorrelation on statistical process control procedures.

5.2 Control Charts for the Unemployment Insurance Data

Figure 6 is a control chart of the residuals ($e_t = z_t - z'_t$) of the ARI (1,1) model identified for the Unemployment Insurance sampling frame data. Since the model diagnostics support the conclusion that the residuals are independent and identically distributed (i.i.d.) $N(0, \sigma_e^2)$, the residuals are standardized, so that the chart's center line is 0 and the control limits are set at ± 3 . The chart includes model residuals for the sampling frame sizes in the 52 week baseline period and subsequent calendar quarter. The difference between the size of the sampling frame for week 56 and the value predicted by the model falls outside the upper control limit, signaling a special cause.

As an alternative to charting the model residuals, control charts for the Unemployment Insurance sampling frame

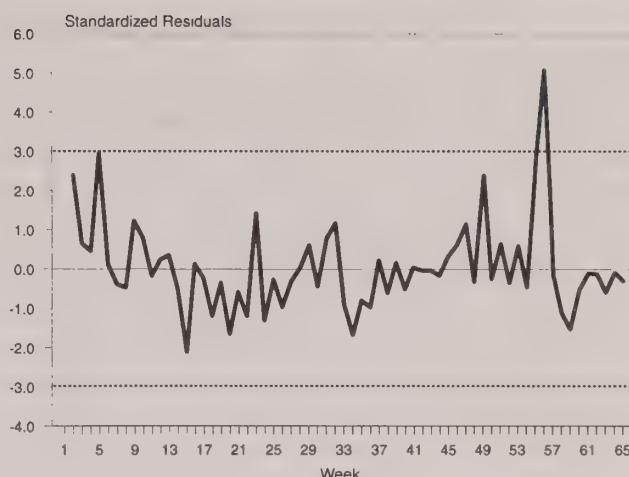


Figure 6. Control chart for model residuals (baseline data + next quarter).

sizes can be constructed. The original observations must be transformed to achieve stationarity, if necessary. The estimated parameters of the time series model are used to construct the mean and control limits of the chart. The variance of an AR(1) process is $\sigma^2 = \sigma_e^2 / (1 - \phi_1^2)$. For the time series model of first differences, ϕ_1' is $-.4045$, and the estimated residual variance, $\sigma_e'^2$, is 184,275,853. The estimated process variance is $184,275,853 / (1 - .1636)$ or 220,325,579.4, and the process standard deviation is 14,843.4. The upper and lower control limits are set at $\pm 3\sigma'$ from the estimated mean difference of zero: $\pm 44,530.2$. The control chart is shown in Figure 7 and signals a special cause for observation 56, like the control chart for the residuals in Figure 6.

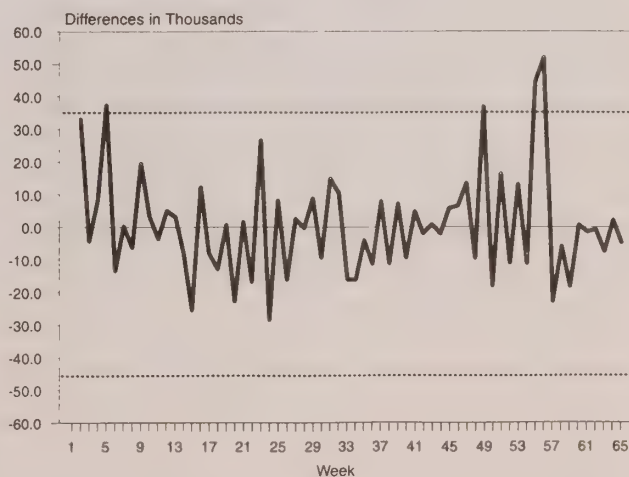


Figure 7. Control chart for UI payments (first differences - baseline + next quarter).

6. CONCLUSIONS

Statistical process control is a useful quality assurance tool for surveys in which samples are selected from frames that are constructed for specified periods from a continuous process. Because the frame sizes constitute a time series, the data may be serially correlated and may have to be transformed in order to achieve stationarity. If the observations are correlated, the appropriate time series (ARIMA) model must be identified in order to estimate the process variance used in setting the control limits. The time series in the preceding example was fitted by a first-order autoregressive integrated (differenced) model – ARI (1,1). More generally, time series may be described by other ARIMA (p, d, q) models, where p is the number of autoregressive terms in the model, d is the degree of differencing to achieve stationarity, and q is the number of moving average terms in the model. Seasonal time series models include additional AR, MA, and differencing parameters for the appropriate lag(s).

Once the model has been identified from baseline data, observations from subsequent periods can be plotted in the control chart. In the control charts in Figures 6 and 7, one calendar quarter (13 weeks) of observations are plotted following the observations from the 52 week baseline. The time series model should be checked periodically, depending on the data collection interval, to determine if the model parameters have changed.

If the statistical process control procedures signal a special cause of variation, survey managers must use other quality management tools to determine the root causes of the frame problems and then implement corrective actions to improve survey procedures. Survey managers can move from troubleshooting and error correction to continuous improvement of the survey process by systematically removing the assignable causes of variation identified through statistical process control.

In the case of the Unemployment Insurance sampling frame data, the special cause was not preventable: the volume of Unemployment Insurance payments spiked during a week which followed a short work week due to a holiday and which coincided with a layoff at a large establishment. The large sampling frame was not the result of a technical problem with the construction of the frame. In other states, at different time periods, statistical process control has detected errors as diverse as data entry mistakes (a frame of 558,432 reported instead of 5,558,432), omission of the Unemployment Insurance transactions for one of five work days, resulting in an approximate 20 percent decrease in the frame size, and the failure to

update edits in the sample selection software, which caused 'foreign elements to enter the frame.

The procedure described in this article is applicable to other areas of survey and information management in addition to the integrity of sampling frames. The procedure can be used to reduce nonsampling error attributable to data recording or data entry for surveys conducted daily, monthly, *etc.* More generally, statistical process control can be used to assure the integrity of databases or management information systems whenever information is collected or reported in subgroups, such as data collected at multiple sites or by several researchers or auditors.

ACKNOWLEDGEMENT

The author wishes to thank the reviewers for their helpful comments and suggestions.

REFERENCES

- ALWAN, L.C., and ROBERTS, H.V. (1988). Time series modeling for statistical process control. *Journal of Business and Economic Statistics*, 6, 87-95.
- BERTHOUEX, P.M., HUNTER, W.G., and PALLESEN, L. (1978). Monitoring sewage treatment plants: some quality control aspects. *Journal of Quality Technology*, 10, 139-149.
- BOX, G.E.P., and PIERCE, D.A. (1970). Distribution of residual autocorrelations in autoregressive moving average time series models. *Journal of the American Statistical Association*, 65, 1509-1526.
- BOX, G.E.P., and JENKINS, G.M. (1976). *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day.
- DEMING, W.E. (1982). *Quality, Productivity, and Competitive Position*. Cambridge: Massachusetts Institute of Technology Center for Advanced Engineering Study.
- MARAGAH, H.D., and WOODALL, W.H. (1992). The effect of autocorrelation on the retrospective X -chart. *Journal of Statistical Computation and Simulation*, 40, 29-42.
- MONTGOMERY, D.C., and MASTRANGELO C.M. (1991). Some statistical process control methods for autocorrelated data. *Journal of Quality Technology*, 23, 179-204.
- RYAN, T.P. (1989). *Statistical Methods for Quality Improvement*. New York: John Wiley and Sons.
- VASILOPOULOS, A.V., and STAMBOULIS, A.P. (1978). Modification of control chart limits in the presence of data correlation. *Journal of Quality Technology*, 10, 20-30.
- WOODALL, W.H., and FALTIN, F.W. (1993). Autocorrelated data and SPC. *American Society for Quality Control Statistics Division Newsletter*, 13, 18-21.

ACKNOWLEDGEMENTS

Survey Methodology wishes to thank the following persons who have served as referees during 1995. An asterisk indicates that the person served more than once.

- C. Alexander, *U.S. Bureau of the Census*
 R. Bell, *The Rand Corporation*
 *D.R. Bellhouse, *University of Western Ontario*
 N. Bennett, *Yale University*
 *D.A. Binder, *Statistics Canada*
 J.-R. Boudreau, *Statistics Canada*
 J. Brehm, *Duke University*
 F.J. Breidt, *Iowa State University*
 S.J. Butani, *U.S. Bureau of Labor Statistics*
 B.D. Causey, *U.S. Bureau of the Census*
 R.L. Chambers, *Australian National University*
 G. Chen, *University of Regina*
 J. Chen, *University of Waterloo*
 G.H. Choudhry, *Statistics Canada*
 M.L. Cohen, *National Academy of Sciences*
 M.J. Colledge, *Australian Bureau of Statistics*
 J.L. Czajka, *Mathematica Policy Research*
 T. DeMaio, *U.S. Bureau of the Census*
 J. Denis, *Statistics Canada*
 J.-C. Deville, *INSEE*
 J.D. Drew, *Statistics Canada*
 J.-J. Droesbeke, *Université Libre de Bruxelles*
 D. Findley, *U.S. Bureau of the Census*
 W.A. Fuller, *Iowa State University*
 J. Gambino, *Statistics Canada*
 M. Gonzalez, *U.S. Office of Management and Budget*
 R.M. Groves, *University of Maryland*
 K.P. Hapuarachchi, *Statistics Canada*
 M.A. Hidirolou, *Statistics Canada*
 D. Hill, *University of Michigan*
 *D. Holt, *Central Statistical Office, U.K.*
 C.T. Ireland, *U.S. National Security Agency*
 S. Itzhaki, *Hebrew University*
 G. Kalton, *Westat, Inc.*
 P.N. Kokic, *Australian Bureau of Agricultural and Resource Economics*
 P.S. Kott, *National Agricultural Statistical Service*
 M. Kovacevic, *Statistics Canada*
 R.A. Kulka, *Research Triangle Institute*
 S. Kumar, *Statistics Canada*
 *N. Laniel, *Statistics Canada*
 M. Latouche, *Statistics Canada*
 *P. Lavallée, *Statistics Canada*
 L. Lazzeroni, *Stanford University*
 *H. Lee, *Statistics Canada*
 N. Luther, *East-West Center*
 *L. Mach, *Statistics Canada*
 T.K. Mak, *Concordia University*
 *H. Mantel, *Statistics Canada*
 A. Mason, *East-West Center*
 P. Merkouris, *Statistics Canada*
 *D. Pfeiffermann, *Hebrew University*
 H. Pold, *Statistics Canada*
 R.F. Potthoff, *Duke University*
 B. Quenneville, *Statistics Canada*
 T.E. Raghunathan, *University of Michigan*
 É. Rancourt, *Statistics Canada*
 *J.N.K. Rao, *Carleton University*
 L.-P. Rivest, *Université Laval*
 K. Rust, *Westat, Inc.*
 I. Sande, *Bell Communications Research, U.S.A.*
 C.-E. Särndal, *Université de Montréal*
 J. Schafer, *Pennsylvania State University*
 *W.L. Schaible, *U.S. Bureau of Labor Statistics*
 *F.J. Scheuren, *George Washington University*
 *J. Sedransk, *State University of New York – Albany*
 R. Sigman, *U.S. Bureau of the Census*
 M. Simard, *Statistics Canada*
 *A. Singh, *Statistics Canada*
 *M.P. Singh, *Statistics Canada*
 R.P. Singh, *U.S. Bureau of the Census*
 *C.J. Skinner, *University of Southampton*
 *D. Stukel, *Statistics Canada*
 *A. Théberge, *Statistics Canada*
 Y. Tillé, *Université Libre de Bruxelles*
 M. Traugott, *University of Michigan*
 J. Trépanier, *Statistics Canada*
 R. Valliant, *U.S. Bureau of Labor Statistics*
 J. Waite, *U.S. Bureau of the Census*
 *J. Waksberg, *Westat, Inc.*
 S. Wing, *Synetics*
 W.E. Winkler, *U.S. Bureau of the Census*
 *K.M. Wolter, *National Opinion Research Center*
 *A. Zaslavsky, *Harvard University*

Acknowledgements are also due to those who assisted during the production of the 1995 issues: S. Beauchamp (Photocomposition), L. Rousseau and S. Cadieux (Official Languages and Translation). Finally we wish to acknowledge S. DiLoreto, S.F. Bertrand, C. Larabie and D. Lemire of Household Survey Methods Division, for their support with coordination, typing and copy editing.

CONTENTS

TABLE DES MATIÈRES

Volume 23, No. 3, September/Septembre 1995

Research papers/Articles

V.P. GODAMBE

- Estimation of parameters in survey sampling: optimality 227

Constance van EEDEN

- Minimax estimation of a lower bounded scale parameter of a gamma distribution for
scale invariant squared error loss 245

Stavros KOUROUKLIS

- Estimation of an exponential quantile under Pitman's measure of closeness 257

Brani VIDAKOVIC and Anirban DasGUPTA

- Lower bounds on Bayes risks for estimating a normal variance: with applications 269

Scott M. JORDAN and K. KRISHNAMOORTHY

- Confidence regions for the common mean vector of several normal populations 283

André Robert DABROWSKI and Abdelhak ZOGLAT

- Strong invariance principles for triangular arrays of weakly dependent random variables 299

Case study in data analysis

Étude de cas en analyse des données

Editor's Introduction

- Effects of growth regulators on silver maple trees: a case study 311

Fernando CAMACHO and Geoffrey ARRON

- Effects of the regulators paclobutrazol and flurprimidol on the growth of terminal
sprouts formed on trimmed silver maple trees 312

Hyun Suk LEE and Bob PHILIPS

- In search of the "best" growth inhibitor 322

Jeff. A. SLOAN, Carl J. SCHWARTZ, and Linda R. NEDEN

- Silver maple trees growth regulators dataset 325

C. SCHWARZ and N. REID

- Comments on the analyses 329

CONTENTS

TABLE DES MATIÈRES

Volume 23, No. 4, December/Décembre 1995

Richard J. COOK and Vern T. FAREWELL Conditional inference for subject-specific and marginal agreement: Two families of agreement measures	333
Mayer ALVO and Paul CABILIO Rank correlation methods for missing data	345
Sneh GULATI and W.J. PADJETT Nonparametric function estimation from inversely sampled record-breaking data	359
Marianthi MARKATOU and Joel L. HOROWITZ Robust scale estimation in the error components model using the empirical characteristic function	369
Gracelia BOENTE and Ricardo FRAIMAN Asymptotic distribution of data-driven smoothers in density and regression estimation under dependence	383
John S.J. HSU Generalized Laplacian approximations in Bayesian inference	399
Alan E. GELFAND and Saurabh MUKHOPADHYAY On nonparametric Bayesian inference for the distribution of a random sample	411
Gilles R. DUCHARME Uniqueness of least distance estimators in regression models with multivariate response	421
Rick ROUTLEDGE and Min TSAO Uniform validity of saddlepoint expansion on compact sets	425

JOURNAL OF OFFICIAL STATISTICS

An International Quarterly Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey Methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents Volume 11, Number 1, 1995

Preface	3
Ten Years of the Journal of Official Statistics	
<i>Fritz Scheuren</i>	5
ISI: Towards the 21st Century	
<i>Zoltan E. Kennessey</i>	11
Challenges Facing the United Kingdom Central Statistical Office	
<i>William McLennan</i>	21
The Statistical Profession and the Chartered Statistician (CStat)	
<i>T.M.F. Smith</i>	33
Planning the Methodology Work Program in a Statistical Agency	
<i>Susan Linacre</i>	41
Methods for Design Effects	
<i>Leslie Kish</i>	55
A Decade of Questions	
<i>Nora Cate Schaeffer</i>	79
Theoretical Motivation for Post-Survey Nonresponse Adjustment in Household Surveys	
<i>Robert M. Groves and Mick P. Couper</i>	93
Controlling Invasion of Privacy in Surveys of Change Over Time – A Non-Technical Review	
<i>Tore Dalenius</i>	107
Changes in Statistical Technology	
<i>Wouter J. Keller</i>	115

Contents Volume 11, Number 2, 1995

Increasing Response to Personally-Delivered Mail-Back Questionnaires <i>Don A. Dillman, Dana E. Dolsen, and Gary E. Machlis</i>	129
Data Quality in a CAPI Survey: Keying Errors <i>Lynn Dielman and Mick P. Couper</i>	141
Understanding the Standardized/Non-Standardized Interviewing Controversy <i>Paul Beatty</i>	147
The Evolution and Development of Agricultural Statistics at the United States Department of Agriculture <i>Frederic A. Vogel</i>	161
Methodological Principles for a Generalized Estimation System at Statistics Canada <i>V. Estevao, M.A. Hidirolou, and C.-E. Särndal</i>	181
An Agenda for Research in Statistical Disclosure Limitation <i>Lawrence H. Cox and Laura V. Zayatz</i>	205
Miscellanea	
The central Register of Population of the Republic of Slovenia <i>Irena Trsinar</i>	221
Book Review	225
In Other Journals	231

Contents Volume 11, Number 3, 1995

Sources of Data on Socio-Economics Differential Mortality in the United States <i>Donna L. Hoyert, Gopal K. Singh, and Harry M. Rosenberg</i>	233
Quantifying Errors in the Swedish Consumer Price Index <i>Jörgen Dalén</i>	261
Estimating Distribution Functions with Auxiliary Information using Poststratification <i>P.L.D. Nascimento and Chris Skinner</i>	277
Weighting Anchors: Verbal and Numeric Labels for Response Scales <i>Colm O'Muircheartaigh</i>	295
Pretesting Procedures at Statistics Sweden's Measurement Evaluation and Development Laboratory <i>Lars R. Bergman</i>	309
Miscellanea	
A Bibliography on Telephone Survey Methodology <i>Anwer Khursid and Hardeo Sahai</i>	325
Special Note	369
In Other Journals	373

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue (Vol. 19, No. 1 and onward) of *Survey Methodology* as a guide and note particularly the following points:

1. Layout

- 1.1 Manuscripts should be typed on white bond paper of standard size (8½ × 11 inch), one side only, entirely double spaced with margins of at least 1½ inches on all sides.
- 1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.
- 1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.
- 1.4 Acknowledgements should appear at the end of the text.
- 1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

- 3.1 Avoid footnotes, abbreviations, and acronyms.
- 3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as “exp(·)” and “log(·)”, etc.
- 3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
- 3.4 Write fractions in the text using a solidus.
- 3.5 Distinguish between ambiguous characters, (e.g., w, ω; o, O; l, 1).
- 3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

4. Figures and Tables

- 4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.
- 4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

5. References

- 5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, p. 164).
- 5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Avant de dactylographier votre texte pour le soumettre, prière d'examiner un numéro récent de *Techniques d'enquête* (à partir du vol. 19, n° 1) et de noter les points suivants:

1. Les textes doivent être dactylographiés sur un papier blanc de format standard (8½ par 11 pouces), sur une face seulement, à double interligne partout et avec des marges d'au moins 1½ pouce tout autour.
- 1.2 Les textes doivent être divisés en sections numérotées portant des titres appropriés.
- 1.3 Le nom et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.
- 1.4 Les remerciements doivent paraître à la fin du texte.
- 1.5 Toute annexe doit suivre les remerciements mais précéder la bibliographie.

2. Résumé

Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.

3. Rédaction

- 3.1 Éviter les notes au bas des pages, les abréviations et les sigles.
- 3.2 Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme exp(·) et log(·) etc.
- 3.3 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin.
- 3.4 Écrire les fractions dans le texte à l'aide d'une barre oblique.
- 3.5 Distinguer clairement les caractères ambigus (comme w, ω; o, O; 1, l).
- 3.6 Les caractères italiques sont utilisés pour faire ressortir des mots. Indiquer ce qui doit être imprimé en italique en le soulignant dans le texte.

4. Figures et tableaux

- 4.1 Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).
- 4.2 Ils doivent paraître sur des pages séparées et porter une indication de l'endroit où ils doivent figurer dans le texte. (Normalement, ils doivent être insérés près du passage qui y fait référence pour la première fois.)

5. Bibliographie

- 5.1 Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence.
Exemple: Cochran (1977, p. 164).
- 5.2 La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.

Contents Volume 11, Number 2, 1995

129	Increasing Response to Personally-Delivered Mail-Back Questionnaires
	<i>Don A. Dillman, Dana E. Dolson, and Gary E. Machlis</i>
141	Data Quality in a CAPI Survey: Keying Errors
	<i>Lynn Dieleman and Mick P. Couper</i>
147	Understanding the Standardized/Non-Standardized Interviewing Controversy
	<i>Paul Beatty</i>
161	The Evolution and Development of Agricultural Statistics at the United States Department of Agriculture
	<i>Frederic A. Vogel</i>
181	Methodological Principles for a Generalized Estimation System at Statistics Canada
	<i>V. Estevao, M.A. Hidiroglou, and C.-E. Särndal</i>
205	An Agenda for Research in Statistical Disclosure Limitation
	<i>Lawrence H. Cox and Laura V. Zayatz</i>

Miscellanea

221	The central Register of Population of the Republic of Slovenia
	<i>Irena Trsinar</i>
225	Book Review
231	In Other Journals

Contents Volume 11, Number 3, 1995

233	Sources of Data on Socio-Economics Differential Mortality in the United States
	<i>Donna L. Hoyert, Gopal K. Singh, and Harry M. Rosenberg</i>
261	Quantifying Errors in the Swedish Consumer Price Index
	<i>Jörgen Dalén</i>
277	Estimating Distribution Functions with Auxiliary Information using Poststratification
	<i>P.L.D. Nascimento and Chris Skinner</i>
295	Weighting Anchors: Verbal and Numeric Labels for Response Scales
	<i>Colm O'Muircheartaigh</i>
309	Pretesting Procedures at Statistics Sweden's Measurement Evaluation and Development Laboratory
	<i>Lars R. Bergman</i>

Miscellanea

325	A Bibliography on Telephone Survey Methodology
	<i>Anwer Khurshid and Hardeo Sahai</i>
369	Special Note
373	In Other Journals

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey Methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents Volume 11, Number 1, 1995

Preface	3
Ten Years of the Journal of Official Statistics	5
<i>Fritz Scheuren</i>	
ISI: Towards the 21st Century	11
<i>Zoltan E. Kennessy</i>	
Challenges Facing the United Kingdom Central Statistical Office	21
<i>William McLennan</i>	
The Statistical Profession and the Chartered Statistician (CStat)	33
<i>T.M.F. Smith</i>	
Planning the Methodology Work Program in a Statistical Agency	41
<i>Susan Linacre</i>	
Methods for Design Effects	55
<i>Leslie Kish</i>	
A Decade of Questions	79
<i>Nora Cate Schaeffer</i>	
Theoretical Motivation for Post-Survey Nonresponse Adjustment in Household Surveys	93
<i>Robert M. Groves and Mick P. Couper</i>	
Controlling Invasion of Privacy in Surveys of Change Over Time – A Non-Technical Review	107
<i>Tore Dalenius</i>	
Changes in Statistical Technology	115
<i>Wouter J. Keller</i>	

CONTENTS TABLE DES MATIÈRES

Volume 23, No. 4, December/Décembre 1995

Richard J. COOK et Vern T. FAREWELL	333
Conditional inference for subject-specific and marginal agreement: Two families of agreement measures	
Mayer ALVO et Paul CABILLO	345
Rank correlation methods for missing data	
Sneh GULATI et W.J. PADJETT	359
Nonparametric function estimation from inversely sampled record-breaking data	
Marianti MARKATOU et Joel L. HOROWITZ	
Robust scale estimation in the error components model using the empirical characteristic function	369
Graciela BOENTE et Ricardo FRAIMAN	
Asymptotic distribution of data-driven smoothers in density and regression estimation under dependence	383
John S.J. HSU	
Generalized Laplacian approximations in Bayesian inference	399
Alan E. GELFAND et Saurabh MUKKHOOPADHYAY	
On nonparametric Bayesian inference for the distribution of a random sample	411
Gilles R. DUCHARME	
Uniqueness of least distance estimators in regression models with multivariate response	421
Rick ROUITLEDGE et Min TSAO	
Uniform validity of saddlepoint expansion on compact sets	425

CONTENTS TABLE DES MATIÈRES

Volume 23, No. 3, September/Septembre 1995

Research papers/Articles

V.P. GODAMBE
Estimation of parameters in survey sampling: optimality 227

Constance van EEDEN
Minimax estimation of a lower bounded scale parameter of a gamma distribution for
scale invariant squared error loss 245

Stavros KOUROUKLIS
Estimation of an exponential quantile under Pitman's measure of closeness 257

Brani VIDAKOVIC et Anirban DasGUPTA
Lower bounds on Bayes risks for estimating a normal variance: with applications 269

Scott M. JORDAN et K. KRISHNAMOORTHY
Confidence regions for the common mean vector of several normal populations 283

André Robert DABROWSKI et Abdelhak ZOGLAT
Strong invariance principles for triangular arrays of weakly dependent random variables 299

Case study in data analysis
Etude de cas en analyse des données

Editor's Introduction
Effects of growth regulators on silver maple trees: a case study 311

Fernando CAMACHO et Geoffrey ARRON
Effects of the regulators paclobutrazol and flurprimidol on the growth of terminal
sprouts formed on trimmed silver maple trees 312

Hyun Suk LEE et Bob PHILIPS
In search of the "best" growth inhibitor 322

Jeff. A. SLOAN, Carl J. SCHWARTZ, et Linda R. NEDEN
Dataset on silver-maple-tree growth regulators 325

C. SCHWARZ et N. REID
Comments on the analyses 329

REMERCIEMENTS

Techniques d'enquête désire remercier les personnes suivantes, qui ont accepté de faire la critique d'un article durant l'année 1995. Un astérisque indique que la personne a participé plus d'une fois.

- C. Alexander, U.S. Bureau of the Census
 R. Bell, The Rand Corporation
 * D.R. Bellhouse, University of Western Ontario
 N. Bennett, Yale University
 * D.A. Binder, Statistique Canada
 J.-R. Boudreau, Statistique Canada
 J. Brehm, Duke University
 F.J. Breidt, Iowa State University
 S.J. Butani, U.S. Bureau of Labor Statistics
 B.D. Causey, U.S. Bureau of the Census
 R.L. Chambers, Australian National University
 G. Chen, University of Regina
 J. Chen, University of Waterloo
 G.H. Choudhry, Statistique Canada
 M.L. Cohen, National Academy of Sciences
 M.J. Colledge, Australian Bureau of Statistics
 J.L. Czajka, Mathematica Policy Research
 T. DeMaio, U.S. Bureau of the Census
 J. Denis, Statistique Canada
 J.-C. Deville, INSEE
 J.D. Drew, Statistique Canada
 J.-J. Droesbeke, Université Libre de Bruxelles
 D. Findley, U.S. Bureau of the Census
 W.A. Fuller, Iowa State University
 J. Gambino, Statistique Canada
 M. Gonzalez, U.S. Office of Management and Budget
 R.M. Groves, University of Maryland
 K.P. Hapuarachchi, Statistique Canada
 M.A. Hidiroglou, Statistique Canada
 D. Hill, University of Michigan
 * D. Holt, Central Statistical Office, U.K.
 C.T. Ireland, U.S. National Security Agency
 S. Itzhaki, Hebrew University
 G. Kalton, Westat, Inc.
 P.N. Kokic, Australian Bureau of Agricultural and Resource Economics
 P.S. Kott, National Agricultural Statistical Service
 M. Kovacevic, Statistique Canada
 R.A. Kulka, Research Triangle Institute
 S. Kumar, Statistique Canada
 * N. Laniel, Statistique Canada
 M. Latouche, Statistique Canada
 * P. Lavallée, Statistique Canada
 L. Lazzeroni, Stanford University
 * H. Lee, Statistique Canada
 N. Luther, East-West Center
 * L. Mach, Statistique Canada
 T.K. Mak, Concordia University
 * H. Mantel, Statistique Canada
 A. Mason, East-West Center
 P. Merkouris, Statistique Canada
 D. Pfeffermann, Hebrew University
 * H. Pold, Statistique Canada
 R.F. Potthoff, Duke University
 B. Quenneville, Statistique Canada
 T.E. Raghunathan, University of Michigan
 E. Rancourt, Statistique Canada
 * J.N.K. Rao, Carleton University
 L.-P. Rivest, Université Laval
 K. Rust, Westat, Inc.
 I. Sande, Bell Communications Research, U.S.A.
 C.-E. Särndal, Université de Montréal
 J. Schafer, Pennsylvania State University
 * W.L. Schabale, U.S. Bureau of Labor Statistics
 * F.J. Schuren, George Washington University
 * J. Sedransk, State University of New York - Albany
 R. Sigman, U.S. Bureau of the Census
 M. Simard, Statistique Canada
 * A. Singh, Statistique Canada
 * M.P. Singh, Statistique Canada
 R.P. Singh, U.S. Bureau of the Census
 * C.J. Skinner, University of Southampton
 * D. Stukel, Statistique Canada
 * A. Thèberge, Statistique Canada
 Y. Tillé, Université Libre de Bruxelles
 M. Traugott, University of Michigan
 J. Trépanier, Statistique Canada
 R. Valliant, U.S. Bureau of Labor Statistics
 J. Waite, U.S. Bureau of the Census
 * J. Waksberg, Westat, Inc.
 S. Wang, Synetics
 W.E. Winkler, U.S. Bureau of the Census
 * K.M. Wolter, National Opinion Research Center
 * A. Zaslavsky, Harvard University

On remercie également ceux qui ont contribué à la production des numéros de la revue pour 1995: S. Beauchamp (photocomposition), L. Roussseau et S. Cadioux (Langues officielles et traduction). Finalement on désire exprimer notre reconnaissance à S. D'Illoredo, S.F. Bertrand, C. Larabie et D. Lemire de la Division des méthodes d'enquêtes-ménages, pour leur apport à la coordination, la dactylographie et la rédaction.

- BOY, G.E.P., et PIERCE, D.A. (1970). Distribution of residual autocorrelations in autoregressive moving average time series models. *Journal of the American Statistical Association*, 65, 1509-1526.
- BOY, G.E.P., et JENKINS, G.M. (1976). *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day.
- DEMING, W.E. (1982). *Quality, Productivity, and Competitive Position*. Cambridge: Massachusetts Institute of Technology Center for Advanced Engineering Study.
- MARAGAH, H.D., et WOODALL, W.H. (1992). The effect of autocorrelation on the retrospective X-chart. *Journal of Statistical Computation and Simulation*, 40, 29-42.
- MONTGOMERY, D.C., et MASTRANGELO C.M. (1991). Some statistical process control methods for autocorrelated data. *Journal of Quality Technology*, 23, 179-204.
- RYAN, T.P. (1989). *Statistical Methods for Quality Improvement*. New York: John Wiley and Sons.
- VASILIOPOULOS, A.V., et STAMBOULIS, A.P. (1978). Modification of control chart limits in the presence of data correlation. *Journal of Quality Technology*, 10, 20-30.
- WOODALL, W.H., et FALTIM, F.W. (1993). Autocorrelated data and SPC. *American Society for Quality Control Statistics Division Newsletter*, 13, 18-21.

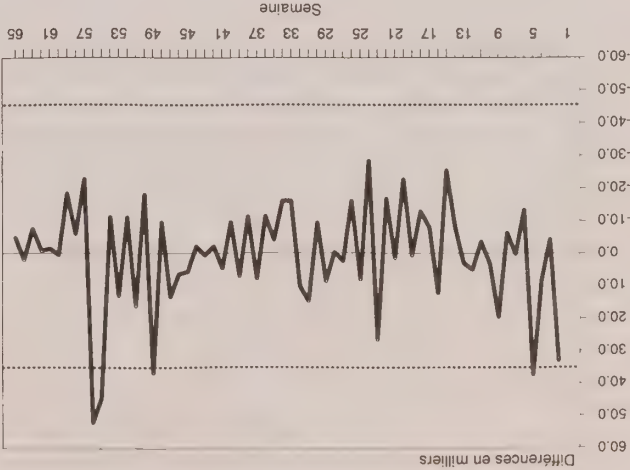


Figure 7. Graphique de contrôle des prestations d'assurance-chômage payées (différences de premier ordre - données de référence + trimestre suivant).

6. CONCLUSIONS

Le contrôle statistique du processus est un instrument de contrôle de la qualité utile pour les enquêtes dont les échantillons sont tirés de bases de sondage construites pour des périodes particulières, par un procédé continu. Comme les tailles des bases de sondage constituent une série chronologique, les données sont parfois liées et doivent alors être transformées pour atteindre la stationnarité. Si les observations sont corrélées, il convient de choisir le modèle chronologique (ARMMI) approprié pour estimer la variance du processus utilisée pour fixer les limites de contrôle. Dans le cas de l'exemple mentionné plus haut, la série chronologique s'accorde avec un modèle autorégressif intégré (à calcul de différences) d'ordre 1 - ARI (1, 1). De façon plus générale, on peut décrire les séries chronologiques au moyen d'autres modèles ARMMI (p, d, q), où p est le nombre de termes autorégressifs dans le modèle, d est le degré de calcul de différences nécessaire pour atteindre la stationnarité, et q est le nombre de termes à moyenne mobile dans le modèle. Les modèles chronologiques saisonniers incluent des paramètres AR, MM et de calcul de différences pour le ou les décalage(s) approprié(s).

Une fois le modèle déterminé à partir des données de référence, on peut reporter sur le graphique de contrôle les observations faites pour les périodes subséquentes. Les graphiques de contrôle des figures 6 et 7 comportent les observations de la période de référence de 52 semaines, et celles faites durant le trimestre suivant (13 semaines). Le modèle chronologique devrait être vérifié périodiquement, en fonction de l'intervalle de collecte des données, afin de s'assurer que les paramètres n'ont pas changé.

Quand les procédures de contrôle statistique du processus signalent l'existence d'une cause spéciale de variation, les chargés d'enquête doivent se servir d'autres instruments

Outre la vérification de l'intégrité des bases de sondage, la méthode décrite dans le présent article s'applique à d'autres domaines de la gestion des enquêtes et de l'information. On peut l'utiliser pour diminuer l'erreur non due à l'échantillonnage attribuable à l'enregistrement ou à la saisie des données dans le cas des enquêtes effectuées quotidiennement, mensuellement, etc. De façon plus générale, le contrôle statistique du processus permet de confirmer l'intégrité des bases de données ou des systèmes de gestion de l'information, dans toutes les situations où l'information est collectée ou enregistrée en sous-groupes, notamment quand les données sont collectées à des sites multiples ou par plusieurs chercheurs ou vérificateurs.

REMERCIEMENTS

L'auteur remercie les réviseurs pour leur suggestions et leurs commentaires judicieux.

BIBLIOGRAPHIE

ALWAN, L.C., et ROBERTS, H.V. (1988). Time series modeling for statistical process control. *Journal of Business and Economic Statistics*, 6, 87-95.

BERTHOUX, P.M., HUNTER, W.G., et PALLESEN, L. (1978). Monitoring sewage treatment plants: some quality control aspects. *Journal of Quality Technology*, 10, 139-149.

5. UTILISATION DU MODÈLE ARMI DANS UN GRAPHIQUE DE CONTRÔLE

5.1 Graphiques de contrôle pour des observations individuelles

On fixe les limites de contrôle pour le graphique d'observations individuelles à $\bar{x} \pm 3\sigma$, où \bar{x} représente la moyenne des observations et σ , l'écart-type estimé du processus. Ryan (1989) examine d'autres méthodes d'estimation de l'écart-type du processus en calculant soit la moyenne des étendues mobiles (la moyenne des écarts absolus entre des observations successives) ou l'écart-type (s) des observations d'échantillon, $\sigma' = s/c$, où c est une constante d'ajustement qui dépend de la taille de l'échantillon. Quand il existe une corrélation sériale des données, l'utilisation de l'écart-type de l'échantillon ou de l'étendue mobile moyenne aboutit parfois à de mauvaises estimations de σ . Les limites de contrôle construites pour ces estimations peuvent donner des résultats très trompeurs, soit parce qu'elles produisent des signaux erronés donnant à penser que le processus est hors de contrôle, ou qu'elles empêchent de déceler certaines causes spéciales de la variation du processus. Le calcul de l'étendue moyenne risque de donner une sous-estimation de σ , car les écarts entre les observations successives sont généralement faibles quand ces dernières sont fortement corrélées. La sous-estimation de σ se traduira par des limites de contrôle trop rapprochées et par une augmentation de la fréquence des signaux indicateurs de causes spéciales. Selon Ryan, quand les données sont corrélées, le calcul de l'écart-type de l'échantillon pour estimer l'écart-type du processus fournit une meilleure estimation de σ que celui de l'étendue mobile moyenne, à condition que l'échantillon comprenne au moins 50 observations. Cependant, l'écart-type de l'échantillon n'est un estimateur non biaisé de σ que quand les observations sont indépendantes.

Vasilopoulos et Stamoulis (1978) ont analysé l'effet de la corrélation sériale des données sur les limites de contrôle des graphiques de \bar{x} et s (écart-type), et développé des équations définissant certains facteurs qui permettent d'ajuster les limites de contrôle des données produites par un processus autorégressif. Autrement, on peut définir un modèle chronologique pour les données corrélées et construire un graphique de contrôle en se servant des résidus du modèle pour surveiller le processus. Cette méthode est décrite par Berthouex, Hunter et Palleisen (1978) pour des sous-groupes de mesures de variables environnementales collectées par des stations d'épuration de l'eau. Alwan et Roberts (1988) se servent des résidus de modèles à moyenne mobile pondérée exponentiellement (MMEP) pour les séries chronologiques tant stationnaires que non stationnaires. Montgomery et Masttragelo (1991) utilisent les résidus d'un modèle autorégressif dans un graphique MMEP et prétendent que ce type de graphique peut servir d'approximation à bon nombre de modèles autocorrélés, particulièrement si la corrélation des observations est positive et que la moyenne ne dévie pas trop rapidement. En outre, le lecteur trouvera dans Maragah et Woodall (1992) et dans Woodall et Faltin (1993) une discussion supplémentaire des effets de l'autocorrélation sur les méthodes de contrôle statistique des processus.

5.2 Graphiques de contrôle des données sur l'assurance-chômage

La figure 6 représente un graphique de contrôle des résidus ($e_t = z_t - z'_t$) du modèle ARI (1,1) choisi pour les données sur les bases de sondage relatives à l'assurance-chômage. Puisque les vérifications du modèle confirment que les résidus sont indépendants et identiquement distribués (i.i.d.) $N(0, \sigma_e^2)$, on a normalisé ces derniers, de sorte que la ligne centrale du graphique corresponde à la valeur 0, et fixe les limites de contrôle à ± 3 . Le graphique inclut les résidus du modèle pour les tailles de la base de sondage observées durant la période de référence de 52 semaines et durant le trimestre suivant. Pour la semaine 56, la différence entre la taille de la base de sondage et la valeur prédite par le modèle tombe au-delà de la limite de contrôle supérieure, ce qui indique l'existence d'une cause spéciale de variation.

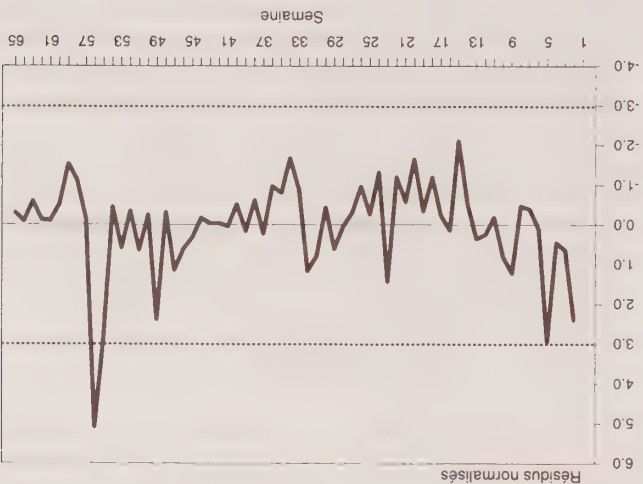
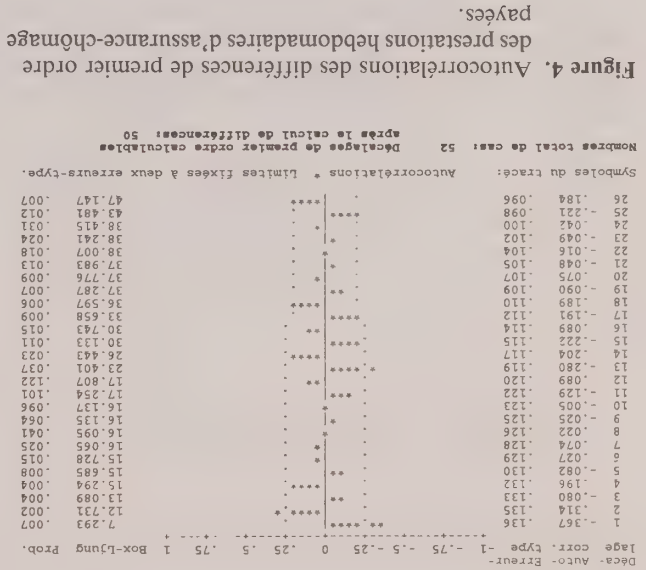


Figure 6. Graphique de contrôle des résidus du modèle (données de référence + trimestre suivant).

Au lieu de représenter graphiquement les résidus du modèle, on peut construire un graphique de contrôle de la taille des bases de sondage. Au besoin, on doit transformer les observations originales pour atteindre la stationnarité. On se sert des estimations des paramètres du modèle chronologique pour déterminer la moyenne et les limites de contrôle du graphique. La variance d'un processus AR(1) est $\sigma^2/(1 - \phi_1^2)$. Pour le modèle chronologique des différences d'ordre 1, ϕ_1^2 est égal à -0.4045 , et l'estimation de la variance résiduelle, σ_e^2 , est $184,275,853$. L'estimation de la variance du processus est $184,275,853/(1 - 0.1636)$, ou $220,325,579.4$, et l'écart-type du processus est égal à $14,843.4$. On fixe les limites de contrôle inférieure et supérieure à $\pm 3\sigma$, soit $\pm 44,530.2$, par rapport à la différence moyenne estimée, égale à zéro. Le graphique de contrôle (figure 7) indique qu'il existe une cause spéciale de variation pour l'observation 56, tout comme le graphique de contrôle des résidus présenté à la figure 6.

$(1 - B)^2 z_t = (z_t - z_{t-1}) - (z_{t-1} - z_{t-2})$, le calcul de différences saisonnières, ou encore, à des transformations logarithmiques ou à d'autres procédés de stabilisation de la variance.

Les autocorrélations des différences d'ordre 1 de la série chronologique, que l'on trouve à la figure 4, sont compatibles avec un processus stationnaire. Les autocorrélations diminuent rapidement, tandis que les autocorrélations partielles (non représentées) disparaissent après le premier décalage. Ces observations donnent à penser qu'on peut modéliser les données au moyen d'un processus autorégressif intégré d'ordre 1, $AR(1,1)$. Le terme d'autorégression AR indique qu'on n'estimera qu'un seul paramètre autorégressif, et le caractère d'intégration (I) montre qu'on a transformé la série chronologique originale en calculant les différences d'ordre 1.



4.4 Estimation du modèle

On a estimé le modèle grâce à la procédure ARMMI du logiciel SPSS Trends (version 4.0), qui est basée sur les travaux de Box et Jenkins.

Le modèle provisoire est:

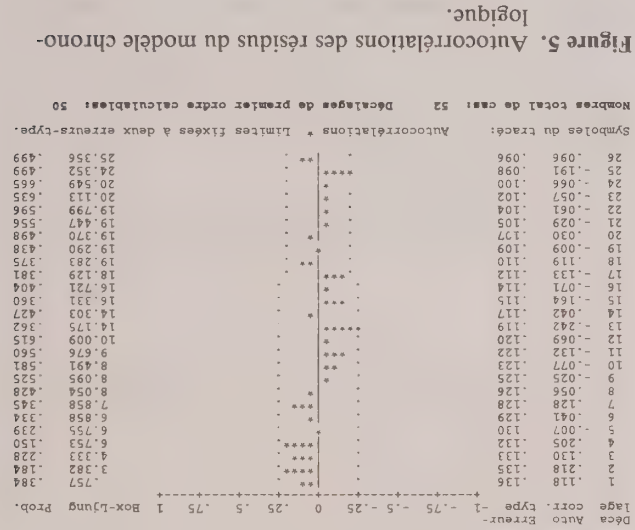
$$z_t' = (1 + \phi_1)z_{t-1} - \phi_1 z_{t-2} + e_t, \text{ ou } z_t - z_{t-1} = \phi_1(z_{t-1} - z_{t-2}) + e_t,$$

où ϕ_1 est le paramètre autorégressif d'ordre 1, et e est le terme d'erreur, qu'on suppose distribué normalement, avec une moyenne égale à 0 et une variance σ_e^2 . L'estimation du paramètre autorégressif, ϕ_1 , est égale à $- .4045$, et l'estimation de la variance résiduelle, σ_e^2 , est $184,275,853$ (avec 50 degrés de liberté). Le signe négatif du paramètre AR est compatible avec les signes alternants des autocorrélations à la figure 4. Le modèle n'inclut pas de

constante, car l'estimation de la moyenne du processus ne diffère pas de zéro de façon significative.

4.5 Vérifications du modèle

On peut déterminer si le modèle estimé convient aux données observées en examinant les résidus. Si le modèle est bien ajusté aux données, les résidus (e_t) devraient être des "bruits blancs", c'est-à-dire non corrélés. La figure 5 montre les autocorrélations des résidus du modèle. Bien que l'autocorrélation pour le décalage 13 soit significative, la valeur de la variable statistique Q de Box-Ljung n'est pas significative jusqu'au décalage 13. (La variable statistique Q teste la signification des autocorrélations pour les décalages 1 à k . Pour une discussion approfondie, consulter Box et Pierce (1970)). En outre, aucune autocorrélation partielle (non représentée dans la figure) n'est significative. Ces résultats indiquent qu'il n'existe pas de corrélation sériale des résidus.



Pour tester l'hypothèse selon laquelle la distribution des résidus du modèle est normale, $N(0, \sigma_e^2)$, on a effectué le test de validité de l'ajustement de Kolmogorov-Smirnov ($K-S$). Pour l'estimation de la variance égale à $184,275,853$, la variable à tester est égale à $.591$ ($p = .876$) et on ne peut donc rejeter l'hypothèse selon laquelle la distribution des différences est normale.

Pour que le processus $AR(1)$ soit stationnaire, la valeur absolue du paramètre autorégressif doit être inférieure à un. Afin de tester l'hypothèse que $|\phi_1| \geq 1$ pour le modèle, on calcule: $t = (|\phi_1| - 1) / SE(\phi_1)$, où $|\phi_1|$ est la valeur absolue de l'estimation du paramètre autorégressif, et $SE(\phi_1)$, l'erreur-type de ϕ_1 . Le modèle statistique donne $t = (.4045 - 1) / .1295$, ou $t = -4.6$. La probabilité que la valeur absolue de ϕ_1 soit aussi petite que $.4045$, si la valeur absolue réelle de $\phi_1 \geq 1$, est très faible (< 0.00001). On rejette donc l'hypothèse que $|\phi_1| \geq 1$, et on conclut que la série des différences d'ordre 1 est stationnaire.

les concepts de stationnarité et d'autocorrélation, afin d'expliquer les méthodes utilisées pour choisir le modèle fondamental de l'analyse des séries chronologiques devraient consulter un des nombreux traités sur le sujet, en particulier Box et Jenkins (1976).

4.1 Stationnarité

On peut imaginer les observations individuelles qui constituent une série chronologique comme une collection de variables aléatoires à plusieurs dimensions $- (z_1, \dots, z_n) -$ où p est une densité de probabilité et où z_1, \dots, z_n sont des variables aléatoires. Si la distribution conjointe des variables aléatoires ne varie pas en fonction du temps, c'est-à-dire si $p(z_1, \dots, z_{t+n}) = p(z_1, \dots, z_t + n + m)$, le processus est dit *strictement* stationnaire. En pratique, il est difficile d'établir une stationnarité stricte. Dans le cas de la présente application, on présume que la série chronologique est *faiblement* stationnaire. On parle aussi de stationnarité d'ordre 2, car les premiers et deuxième moments du processus sont invariants en fonction du temps $- E(z_t) = E(z_{t+m}), VAR(z_t) = VAR(z_{t+m})$, et $COV(z_t, z_{t+k}) = COV(z_{t+m}, z_{t+k+m})$.

Jusqu'à la fin du présent article, les termes *stationnaire* ou *stationnarité* feront allusion à un processus qui répond aux critères de faible stationnarité.

4.2 Autocorrélation

Dans le cas d'une série chronologique stationnaire, la covariance entre deux observations dépend uniquement du nombre de périodes (décalages) qui les séparent $- COV(z_t, z_{t+k}) = COV(z_{t+m}, z_{t+k+m})$. La corrélation de z_t et de z_{t+k} est égale à $COV(z_t, z_{t+k}) / VAR(z_t)$ et est représentée par la notation ρ_k , où k correspond au nombre de périodes entre les observations. Par exemple, ρ_1 représente la corrélation des observations de la série chronologique séparées par une période, et est égal à $COV(z_t, z_{t+1}) / VAR(z_t)$. Une corrélation pour la période k reçoit le nom d'autocorrélation, parce qu'il s'agit de la corrélation entre des observations qui forment une série chronologique. Les autocorrélations pour les divers décalages peuvent être représentées sur un graphique, appelé *corrélogramme*, qui facilite le choix du modèle approprié pour une série chronologique.

4.3 Choix du modèle chronologique

La figure 2 représente le corrélogramme de la série chronologique, couvrant 52 semaines, du nombre de prestations hebdomadaires d'assurance-chômage payées dans les bases de sondage. Les autocorrélations diminuent ou "disparaissent" très lentement, phénomène qui est caractéristique d'un processus non stationnaire. (À nouveau, le lecteur trouvera dans Box (1976) et dans d'autres traités sur les séries chronologiques une discussion approfondie du choix du modèle.)

Le calcul de différences est une méthode qui permet de transformer une série non stationnaire en une série

stationnaire. Le symbole B est l'opérateur de décalage qui, appliqué à z_t , fait reculer l'indice inférieur d'une période. Donc, la différence d'ordre 1 de z_t est $(1 - B)z_t = z_t - z_{t-1}$.

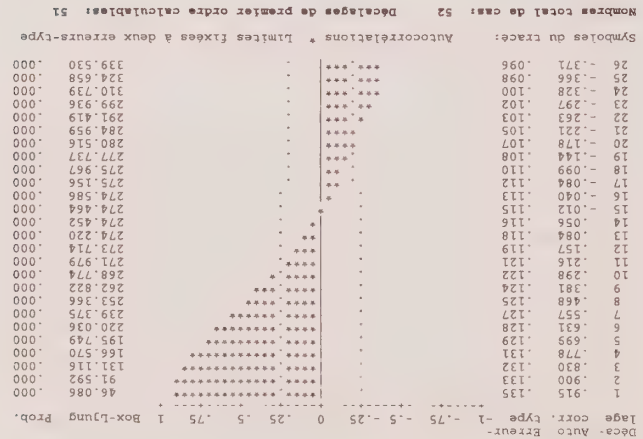


Figure 2. Autocorrélation de la série chronologique des prestations hebdomadaires d'assurance-chômage payées.



La figure 3 représente la série chronologique des différences $z_t - z_{t-1}$ pour les données sur les bases de sondage de l'assurance-chômage. Cette série paraît stationnaire, autour d'une moyenne égale à zéro. (L'estimation de la moyenne d'échantillon pour les différences est 150.8, avec une erreur-type de 2,064.0. La variable à tester $t = (150.8 - 0) / 2.064$ est égale à .07, et on ne peut donc rejeter l'hypothèse selon laquelle $\mu = 0$). Pour d'autres séries chronologiques, le calcul de différences d'ordre 1 pourrait ne pas permettre d'atteindre la stationnarité. Le cas échéant, il est nécessaire de recourir à des transformations telles que le calcul de différences d'ordre 2 -

4. ANALYSE DES DONNÉES ET DÉVELOPPEMENT

DU MODÈLE

La figure 1 représente un graphique chronologique de la taille des bases de sondage pendant une période de 52 semaines. La base de sondage de chaque semaine comprend les bénéficiaires de l'assurance-chômage de la semaine précédente qui continuent à recevoir une prestation, moins les bénéficiaires de l'assurance-chômage de la semaine précédente qui sont retournés au travail, n'ont plus droit à une prestation ou ont omis de présenter une demande, plus les requérants nouvellement admissibles ainsi que les requérants admissibles qui n'ont pas présenté de demande ou reçu de prestation pour une demande la semaine précédente.

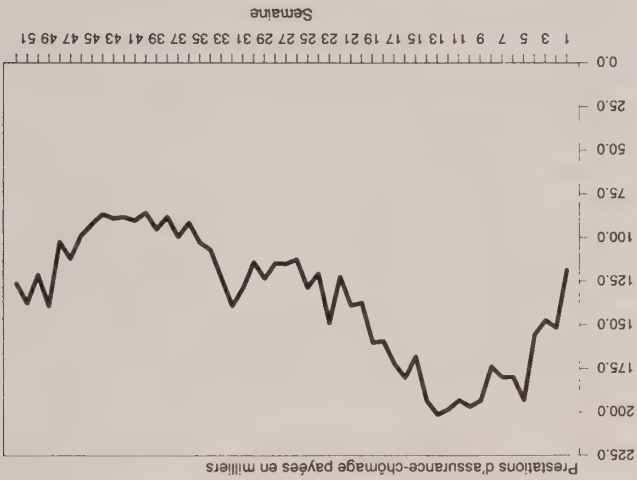


Figure 1. Nombre de prestations d'assurance-chômage payées par semaine.

On établit les graphiques de contrôle pour les observations individuelles en supposant que les données sont indépendantes et identiquement distribuées (i.i.d.). Toutefois, l'existence d'une corrélation sériale des données risque de fausser considérablement les estimations de la variance du processus (et, donc, les limites de contrôle). Aussi doit-on déterminer si les observations sont liées avant de pouvoir construire les graphiques de contrôle relatifs aux bases de sondage de l'assurance-chômage.

Le graphique de la série chronologique présenté à la figure 1 fournit une preuve visuelle que les observations ne sont pas indépendantes. Les données sur les bases de sondage affichent des tendances distinctes, caractérisées par l'augmentation des valeurs durant les 13 semaines du premier trimestre, puis leur diminution au cours des deux trimestres suivants, et enfin leur augmentation durant les 13 semaines du dernier trimestre. On peut tester l'existence de la corrélation sériale évoquée par la représentation graphique de la figure 1 au moyen des méthodes établies pour analyser les séries chronologiques. Bien qu'une discussion approfondie de l'analyse des séries chronologiques dépasse le cadre du présent article, on examinera

3.2 Méthodes d'échantillonnage relatives au contrôle de la qualité des prestations et sources d'erreur

Chaque État tire hebdomadairement des échantillons aléatoires de versements au titre de l'assurance-chômage qu'on examine afin de déterminer si le requérant a reçu le bon montant. Si ce dernier est incorrect, l'enquêteur détermine les types d'erreurs et leurs causes, afin que les gestionnaires du programme puissent prendre des mesures correctives. Les bases de sondage sont constituées chaque semaine en se fondant sur l'ensemble des prestations d'assurance-chômage versées entre le dimanche à 12 h et le samedi suivant à 23 h 59. Un programme informatique permet de vérifier la base de données de l'État pour s'assurer que seuls les versements conformes à la définition opérationnelle de la population cible du programme soient inclus dans la base de sondage. Par exemple, on exclut de celle-ci les paiements effectués au titre de certains programmes d'assurance-chômage temporaires ou peu importants.

Le nombre de chèques d'assurance-chômage émis chaque semaine (et, par conséquent, la taille de la base de sondage) varie selon le nombre de personnes qui demandent des prestations et reçoivent ces dernières durant la semaine en question. Cependant, il existe plusieurs sources d'erreurs potentielles, susceptibles d'avoir une incidence sur l'inité- grité de la base de sondage. Voici quelques-unes des erreurs les plus graves.

Les paiements faits par certains bureaux locaux d'assurance-chômage risquent de ne pas être repris dans la base de données centrale de l'État, en raison de problèmes de télé- communication ou de TAD.

Si l'État crée un fichier distinct pour chaque journée de transaction, les transactions d'un ou de plusieurs jours peuvent être omises par erreur dans le fichier cumulatif final.

Le codage incorrect des transactions peut entraîner soit l'inclusion d'éléments étrangers à la base de sondage, ou l'élimination de transactions qui devraient y être incluses.

Contrôle statistique du processus relatif aux bases de sondage

A.W. SPISAK¹

RÉSUMÉ

Le contrôle statistique du processus est un instrument qui peut servir à vérifier l'exactitude des bases de sondage construites périodiquement. La taille des bases de sondage est reportée sur un graphique de contrôle afin de détecter les causes spéciales de variation. On décrit les méthodes qui permettent de choisir le modèle chronologique (ARMI) approprié pour des observations liées. On examine aussi les applications de l'analyse des séries chronologiques à la construction de graphiques de contrôle. Les données du programme de contrôle de la qualité des prestations d'assurance-chômage du United States Department of Labor servent à illustrer la méthode.

MOTS CLÉS: Autocorrélation; modèles ARMI; graphiques de contrôle; contrôle de la qualité.

1. INTRODUCTION

L'intégrité de la base de sondage est d'une importance capitale dans le domaine de la recherche par sondage. Les imperfections de la base de sondage incluent les éléments manquants (base incomplète), les agrégats d'éléments (plus d'un élément sur une seule liste), les blancs ou les éléments étrangers, et les listes répétées. Ces imperfections peuvent causer plusieurs difficultés, car elles contribuent à l'erreur non due à l'échantillonnage, limitent le nombre d'unités d'échantillonnage tirées des sous-classes de la population, et obligent à utiliser des coefficients de pondération complexes pour estimer les caractéristiques de la population. Les méthodes qui visent à réduire au minimum les problèmes liés aux bases de sondage, ou à limiter leurs répercussions sur l'enquête, sont exposées en détail dans la plupart des traités sur les enquêtes statistiques.

Le présent article met l'accent sur le contrôle statistique du processus relatif aux bases de sondage constituées périodiquement (quotidiennement, hebdomadairement ou mensuellement, par exemple) et constituées d'éléments générés selon un processus continu. Etant donné les fluctuations inhérentes à tout processus dynamique, la taille des bases de sondage varie inévitablement. Or, comment sait-on si les modifications de la taille des bases de sondage reflètent les variations aléatoires du processus, plutôt que des erreurs de construction? Le contrôle statistique du processus permet aux chargés d'enquête de faire la distinction entre les variations inhérentes au processus (causes ordinaires) et celles qui sont indicatrices d'un problème éventuel de construction de la base de sondage (causes spéciales).

2. VARIATIONS DU PROCESSUS ET CONTRÔLE STATISTIQUE DU PROCESSUS

Ces dernières années, les gestionnaires des secteurs privés de la fabrication et des services et ceux du secteur public de l'économie ont été de plus en plus nombreux à adopter les théories sur le contrôle de la qualité proposées par

W. Edwards Deming, J.M. Juran, Philip B. Crosby, Kaoru Ishikawa et d'autres. La gestion de la qualité englobe tout un éventail d'instruments et de méthodes, dont le recours aux graphiques de contrôle pour déterminer si un processus est sous contrôle statistique. Selon Deming (1982), on arrive à ce dernier en éliminant les causes spéciales de variation, de sorte que seules subsistent les variations aléatoires d'un processus stable. L'évolution d'un processus sous contrôle statistique est prévisible.

Distinguer entre les causes ordinaires et spéciales de variation est un principe essentiel du contrôle statistique des processus. Deming (1982) attribue à Walter A. Shewhart, qui a défini nombre des principes du contrôle statistique durant les années 20 et 30, l'énonciation du concept des causes spéciales ou assignables. Les causes spéciales sont ordinairement attribuables à un élément du processus, comme un travailleur, une machine ou un bureau, et se manifestent répétitivement à moins d'être repérées et éliminées. Elles sont signalées par des points de données qui tombent en dehors des limites de contrôle, par des points consécutifs situés au-dessus ou au-dessous de la moyenne, ou par des séries d'observations de valeurs croissantes ou décroissantes. Les causes ordinaires de variation, quant à elles, sont inhérentes au processus. Présentes en tout temps, elles ont une incidence sur l'ensemble du processus. On les limite ou on les élimine grâce à des mesures de gestion visant à modifier ce dernier.

3. APPLICATION DU CONTRÔLE STATISTIQUE DU PROCESSUS À LA CONSTRUCTION DE BASES DE SONDAJE POUR LES ENQUÊTES PÉRIODIQUES

3.1 Programme américain de contrôle de la qualité des prestations d'assurance-chômage

Le programme de contrôle de la qualité des prestations d'assurance-chômage du United States Department of Labor est un exemple de l'utilisation du contrôle statistique

¹ A.W. Spisak, Mathematical Statistician, Unemployment Insurance Service, U.S. Department of Labor, Washington, DC 20210, U.S.A.

- ENQUÊTE MONDIALE SUR LA FÉCONDITÉ (1986). Rapport final. Institut international de statistique, Pays-Bas.
- FERRY, B., et CANTREILLE, P. (1988). L'utilisation de micro-ordinateurs de terrain pour la collecte en démographie. Dans *Congrès africain sur la population*, Dakar, Sénégal, Liege, Belgique: International Union for the Scientific Study of Population (IUSSP), 15-30.
- FORSTER, D., BEHRENS, R.H., CAMPBELL, H., et BYASS, P. (1991). Evaluation of a computerized field data collection system for health surveys. *Bulletin of the World Health Organisation*, 69, 107-111.
- FORSTER, D., et SNOW, R.W. (1992). Using microcomputers for rapid data collection in developing countries. *Health Policy and Planning*, 7, 667-71.
- LYBERG, L. (1985). Plans for computer-assisted data collection at Statistics Sweden. *Bulletin de l'Institut International de Statistique*, actes de la 45^{ème} session, communications demandées, tome L1, livre 3, section 18.2.
- NICHOLLS, W.L., et GROVES, R.M. (1986). The status of computer-assisted telephone interviewing: Part I – Introduction and impact on cost and timeliness of survey data. *Journal of Official Statistics*, 2, 93-115.
- REITMAIER, P., DUPRET, A., et CUTTING, W.A.M. (1987). Better health data with a portable microcomputer at the periphery: An anthropometric survey in Cape Verde. *Bulletin de l'Organisation mondiale de la santé*, 65, 651-657.
- SNOW, R.W., MUNG'ALA, V.O., FORSTER, D., et MARSH, K. (1994). The role of the district hospital in child survival on the Kenyan Coast. *African Journal of Health Sciences*, 1, 71-75.
- TIMAEUS, I.M. (1991). Measurement of adult mortality in less developed countries: a comparative review. *Population Index*, 57, 552-568.
- VLASSOFF, C., et TANNER, M. (1992). The relevance of rapid assessment to health research and interventions. *Health Policy and Planning*, 7, 1-9.

5. DISCUSSION

considérable. Le recours à un tel système de collecte devrait réduire les retards inacceptables dans la présence des données observées lors d'enquêtes telles que l'Enquête mondiale sur la fécondité (tableau 1). La présente étude comparative s'est déroulée dans un contexte démographiques où le personnel recruté n'est pas familier avec l'administration des questionnaires. Nous croyons donc que les résultats dont nous faisons état représentent l'amélioration minimale à laquelle il est permis de s'attendre dans la qualité des données. Le coût initial de la mise en place d'un tel système d'enquêtes est peut-être impressionnant, mais il pourra être amorti si on peut l'utiliser pour des enquêtes répétées ou pour une série d'enquêtes de diverses natures.

REMERCIEMENTS

Les auteurs tiennent à remercier toutes les personnes qui ont participé à l'enquête et notamment les deux chefs d'équipe, Rodgers Chisengwa et Lewis Mitsanze, et les préposés à la saisie, Robert Mutai et Monica Omondi. Merci également à M. Ian Timaeus, qui a conçu le questionnaire sur la mortalité des adultes, et à M. Chris Nevill, qui a réalisé le nouveau recensement. Nous remercions M. Kevin Marsh pour son aide dans la réalisation des enquêtes assistées par ordinateur à Kilifi, le Wellcome Trust (Royaume-Uni) pour son aide financière, la société Psion PLC (Royaume-Uni) qui nous a donné un ordinateur de poche Psion Series 3 et le directeur du KEMRI qui nous a autorisés à publier ses résultats. Les travaux de M. Bob Snow sont financés en partie par le programme de bourses pour la recherche fondamentale en sciences biomédicales Wellcome Trust Senior Fellowship.

BIBLIOGRAPHIE

ANKER, M. (1991). Epidemiological and statistical methods for rapid health assessment. *World Health Statistics Quarterly*, 44, 94-97.

BENCH, J., CLARK, C., DUFOUR, J., et KAUSHAL, R. (1994). Computer-assisted interviewing for the labour force survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*.

DHS, KENYA (1989). Report on Kenyan Demographic and Health Survey. Institute for Resource Development/Macro Systems Inc., Columbia, Maryland, USA.

DENTENEER, D., BETHLEHEM, J.G., HUNDEPOOL, A.J., et KEILLER, W.J. (1987). The BLAISE System for Computer-Assisted Processing, Automation in Survey Processing. Netherlands Bureau of Statistics, CBS Select No. 4, 67-76.

Le taux d'erreurs le plus bas enregistré avec un questionnaire classique utilisé par des intervieweurs expérimentés justifiant de 5 années d'expérience dans la collecte des données était en moyenne de 0.1 l'erreur par questionnaire comportant 17 champs. Ces erreurs ont été pratiquement éliminées à l'aide du logiciel d'enquête mis au point pour un ordinateur de poche Psion Series 3. Cette méthode a éliminé la plupart des erreurs de cheminement commises par les intervieweurs dans l'administration du questionnaire (tableau 4) grâce à des modules préétablis de branchement par saut qui ont réduit le taux d'erreurs d'au moins 90%. Si on ajoutait une liste préalable de réponses dans le logiciel, on pourrait par surcroît réduire le taux d'erreurs d'identification des répondants.

Les chefs d'équipe étaient heureux d'utiliser l'ordinateur; ils ont maîtrisé le clavier QWERTY et appris le fonctionnement du logiciel assez rapidement pour être en mesure de se débrouiller seuls au bout de deux jours. Même si aucune enquête officielle n'a été réalisée pour évaluer et quantifier les réactions des personnes interviewées à l'utilisation du Psion, le nombre de commentaires recueillis a été remarquablement faible et personne n'a refusé d'être interviewé.

Le traitement des données de l'enquête complète a employé deux préposés à la saisie qui ont travaillé à temps plein sur deux ordinateurs IBM pendant 92 heures. Un gestionnaire était sur place pour offrir son aide au besoin et déterminer le format de saisie des données. Il convient de souligner que les deux préposés à la saisie connaissaient bien les procédures de saisie des données ainsi que le matériel et le logiciel utilisés. Dans d'autres situations, il faudrait prévoir une surveillance plus étroite et les services d'un gestionnaire des données, ce qui augmenterait les coûts de l'opération. Un système informatisé de collecte des données Psion permettrait au gestionnaire de passer beaucoup moins de temps à transférer les données recueillies chaque jour, ce qui réduirait cet élément du coût du personnel. Néanmoins, le coût initial des ordinateurs Psion Series 3 risque d'être prohibitif, comparativement au coût prévisible du papier et de la reproduction des questionnaires, si on n'envisage pas de les utiliser dans le cadre d'activités futures de collecte des données.

QUESTOR (Ferry et Cantrelle 1988) offre un logiciel propice à la réalisation d'entrevues assistées par ordinateur. Toutefois, ce logiciel nécessite un ordinateur personnel beaucoup plus cher que l'ordinateur de poche Psion. Nous avons été à même de confirmer qu'il sera profitable de poursuivre la mise au point d'un ensemble d'outils appropriés intégrant la technologie des ordinateurs de poche compatibles avec les ordinateurs personnels et qui seront à la fois plus faciles à utiliser sur le terrain, plus robustes et moins énergivores.

Il faut toujours chercher un compromis entre la réduction des taux d'erreurs, le temps et le coût d'une enquête. Le recours aux entrevues assistées par ordinateur peut permettre de réduire à la fois les taux d'erreurs et le temps consacré à la préparation des données et ce, par une marge

4.2 Erreurs

Les erreurs commises ont été comptabilisées sur deux périodes de deux semaines chacune, pour évaluer l'incidence de la familiarisation graduelle des employés avec leur travail. Les 22 intervieweurs (à l'exclusion des deux chefs d'équipe) ont commis 1,704 erreurs sur 1,427 questionnaires au cours de la première période, et 1,049 erreurs sur 1,158 questionnaires au cours de la seconde période. Le taux moyen d'erreurs par questionnaire a donc atteint 1,19 au cours de la première quinzaine, et 0,90 au cours de la seconde. En outre, sur l'ensemble de la période, il a fallu retourner 37 questionnaires (1,2% de toutes les entrevues) pour les faire reprendre parce que les erreurs qu'ils contenaient ne pouvaient être corrigées au bureau. Ces questionnaires contenaient entre 1 et 6 erreurs chacun, pour un total de 61. C'est la question 5 qui a donné lieu au plus grand nombre d'erreurs (17), suivie par la question 6b (15). La compilation des erreurs commises à chaque question est présentée au tableau 4. Quatorze des 22 intervieweurs ont dû reprendre au moins un questionnaire. L'un d'eux a dû en reprendre 8.

Tableau 4
Compilation des erreurs commises par les 22 intervieweurs utilisant des questionnaires-papier (pour le libellé des question, voir figure 1)

	Période 1 (première quinzaine)	Période 2 (seconde quinzaine)
Identification	163	48
Question 1	6	1
Question 2	8	2
Question 3	125	92
Question 4a	201	138
Question 4b	151	93
Question 5	105	61
Question 6a	94	57
Question 6b	65	41
Question 7	14	0
Question 8	109	63
Question 9	51	10
Question 10	178	134
Question 11	13	1
Question 12	108	71
Question 13	53	3
Question 14	204	149
Question 15	19	76
Code de l'intervieweur	37	9
Total des erreurs	1,704	1,049
Total des questionnaires	1,427	1,158

4.3 Coût

Nous présentons au tableau 5 les coûts comparatifs d'une enquête de cette envergure effectuée à l'aide de l'ordinateur Psion et de questionnaires-papier. Les prix indiqués pour les ordinateurs de poche sont les prix de détail recommandés. Le jeu de la concurrence entre les détaillants pourrait entraîner une réduction des prix de ces ordinateurs atteignant 20% des prix indiqués dans ce tableau. Les prix du matériel informatique sont aussi en baisse. Compte tenu des prix en vigueur actuellement, le coût d'achat d'un système Psion pourrait être amorti après 12 à 15 enquêtes portant sur environ 7,000 répondants.

Etude comparative de l'enquête assistée par ordinateur et de la méthode classique avec questionnaire-papier (coûts en £ sterling du Royaume-Uni)		Coût
Equipement requis		
20 ordinateurs Psion Series 3		2,539.00
20 dispositifs de stockage de 1 Mo		2,039.00
Enquête assistée par ordinateur		59.45
80 piles rechargeables		146.20
1 chargeur à piles		15.95
Coût total		4,799.59
14 rames de papier pour 7,000 entrevues		42.00
Reproduction de 7,000 questionnaires		70.00
20 stylos, gommes et liquide correcteur		27.40
20 planchettes à pince		100.00
2 préposés à la saisie des données		70.00
(2 semaines)		
2 surveillants* (un mois plus		85.00
surtemps)		
Coût total		394.40

* Nécessaires pour la vérification manuelle des formulaires à la fin de chaque journée d'entrevues.

4. RÉSULTATS DES TESTS

4.1 Temps

Les entrevues effectuées avec les questionnaires-papier ont duré en moyenne 5.1 minutes, comparativement à 5.0 minutes pour celles réalisées à l'aide de l'ordinateur de poche; les deux méthodes n'étaient donc pas différentes sur ce plan (tableau 2; noter que 215 des 234 entrevues ont été chronométrées). La durée des entrevues variait considérablement (de 1 à 18 minutes); elle dépendait non seulement du nombre de questions sautées, mais également de la clarté et de la cohérence des réponses fournies.

Tableau 2

Comparaison de la méthode d'enquête classique (questionnaire-papier) et de la méthode d'enquête assistée par ordinateur (Pson Series 3)					
Durée de l'entrevue en minutes					
Moyenne	Equipe A	Equipe B	Maximum général	Maximum général	Moyenne générale
Papier	1	16	5.1 (215)*	5.2 (128)	4.9 (87)
Ordinateur	1	18	5.0 (363)	5.5 (190)	4.5 (173)
* Nombre d'entrevues chronométrées.					

Les chefs d'équipe ont pour leur part consacré 2 à 3 heures par jour à la vérification des questionnaires. Par ailleurs, il a fallu en moyenne à chaque préposé 3 heures et 40 minutes pour saisir les données de 500 dossiers (nombre approximatif d'entrevues réalisées par semaine); l'entrée des données en double a donc demandé 7 heures et 20 minutes (tableau 3). La vérification de ces questionnaires a demandé 2 heures et 23 minutes de plus. Les dossiers complétés ont été révisés deux fois pour entrer les corrections, puis vérifiés à nouveau; cette dernière opération a demandé en moyenne 2 heures et 30 minutes pour 500 dossiers.

Tableau 3

Temps nécessaire au traitement des données de 500 questionnaires					
Activité					
Temps moyen					
Vérification des données	4 heures 8 minutes				
Première saisie des données	3 heures 40 minutes				
Seconde saisie des données	3 heures 40 minutes				
Vérification	2 heures 23 minutes				
Révision	2 heures 33 minutes				
Total	18 heures 24 minutes				

au point pour la vérification de la concordance des données entrées à l'aide du logiciel FoxPro décrits plus haut. Les données entrées sur le Pson sont stockées dans un fichier séparé, à raison d'une ligne par entrevue.

Pour être indiquée correctement, la question doit inclure un numéro, le texte de la question et la réponse (choix d'options, caractère ou numéro). La définition devrait en outre préciser à quelle position sur la ligne l'entrée correspondante doit être stockée et quelle est sa longueur. Les réponses numériques peuvent également inclure un nombre préétabli de signes décimaux. La gamme d'entrées acceptables est une caractéristique optionnelle pour les réponses utilisant des chiffres ou des lettres; elle précisera un minimum, un maximum ou les deux à la fois. Les listes d'options peuvent servir à préciser les codes et leurs valeurs.

Les commandes peuvent être intégrées dans le texte des questions afin de permettre l'évaluation au moment de l'administration du questionnaire. Par exemple, la question de contre-vérification finale de la figure 1 requiert une addition. La syntaxe permet d'inclure cette instruction dans le corps du texte de la question. D'autres commandes peuvent contenir des instructions pour sauter une question, effectuer une contre-vérification entre deux réponses ou passer à une question différente.

Le logiciel utilise ainsi une méthode souple et d'application générale de définition du questionnaire. Il incorpore la manipulation des informations saisies et intègre les fonctions arithmétiques dans les questions ou les lignes de commande. La prochaine étape consisterait à élaborer une interface pour la détermination des caractéristiques du questionnaire qui nous éviterait d'avoir à élaborer une définition du questionnaire correcte au plan syntaxique. On peut se procurer ce logiciel auprès des auteurs.

3.2 Conception du test

On a organisé une journée supplémentaire de formation à l'utilisation du Pson pour les deux chefs d'équipe. Cette formation comportait une explication du matériel et du logiciel ainsi que des séances d'exercice en conditions réelles. Ni l'un ni l'autre des deux chefs d'équipe ne possédait une expérience préalable en informatique. Les deux ont effectué des entrevues en alternant, chaque jour, l'utilisation du Pson et des questionnaires-papier. Ces entrevues ont servi de base à la comparaison des deux méthodes.

Les erreurs commises par les 22 intervieweurs utilisant le questionnaire-papier ont été comptées et totalisées aux fins de l'estimation du taux d'erreurs inhérent à cette méthode de collecte des données. Le temps consacré à la vérification des formulaires remplis, à la saisie, à la vérification et à la correction des données après les contrôles de vraisemblance et de concordance a été mesuré tout au long de l'enquête. Une évaluation semblable du temps d'exécution a été faite dans le cas de la collecte assistée par l'ordinateur Pson.

Figure 1. Questionnaire sur la mortalité des adultes

Noms _____

Dates _____

ID _____ - _____ - _____

(pour toutes les femmes âgées de 25 à 44 ans)

J'aimerais vous poser quelques questions sur vos parents naturels ainsi que sur vos frères et sœurs qui ont la même mère que vous.

1. Votre mère est-elle vivante? (1 = oui, 2 = non) []
2. Votre père est-il vivant? (1 = oui, 2 = non) []
3. Avez-vous déjà accouché? (1 = oui, 2 = non) []
- INTERVIEWEUR: Si la répondante n'a jamais accouché, (Q3 = 2), passer à la question 6.
4. Est-ce que (MENTIONNER TOUS LES PARENTS QUI NE SONT PLUS VIVANTS) []
 était toujours vivant lors de la naissance de votre premier enfant?
- Mère de la répondante []
 Père de la répondante []
5. En quelle année votre premier enfant est-il né? []
 Mère de la répondante []
 Père de la répondante []
6. En quelle année (MENTIONNER TOUS LES PARENTS QUI NE SONT PLUS VIVANTS) []
 est-il mort?
7. Combien avez-vous de sœurs nées de votre mère? []
 (TOUJOURS VIVANTES)
- INTERVIEWEUR: S'il n'y a pas de sœur vivante, (Q7 = 0), passer à la question 9.
8. Combien de ces sœurs toujours vivantes ont moins de 15 ans? []
9. Combien de vos sœurs, nées de votre mère, sont maintenant décédées? []
- INTERVIEWEUR: Si aucune sœur n'est décédée, (Q9 = 0), passer à la question 11.
10. Combien de ces sœurs aujourd'hui décédées sont mortes avant d'avoir 15 ans? []
11. Combien avez-vous de frères nés de votre mère? (TOUJOURS VIVANTS) []
- INTERVIEWEUR: S'il n'y a pas de frère vivant, (Q11 = 0), passer à la question 13.
12. Combien de ces frères toujours vivants ont moins de 15 ans? []
13. Combien de vos frères, nés de votre mère, sont maintenant décédés? []
- INTERVIEWEUR: Si aucun frère n'est décédé, (Q13 = 0), passer à la question 15.
14. Combien de ces frères aujourd'hui décédés sont morts avant d'avoir 15 ans? []
- INTERVIEWEUR: Faire la somme des questions 7, 9, 11 et 13: []
 Q7 = []
 Q9 = []
 Q11 = []
 Q13 = []
 = []
15. Je veux m'assurer que j'ai bien compris. Exception faite de vous-même, votre mère a eu _____ []
 enfants au total. Est-ce bien exact?
- INTERVIEWEUR: En cas de discordance, reprendre les questions 7 à 14 et corriger le cas échéant.
- INTERVIEWEUR: Prière de remercier la répondante de sa coopération.

Code de l'intervieweur []

Les enquêtes de cette ampleur comportent de nombreuses étapes de vérification et de codage des données brutes, qui constituent une autre source de retards. Or, comme la rapidité d'exécution est essentielle à l'évaluation de l'état de santé des populations (Anker 1991; Vlassoff et Tanner 1992), la réduction du temps consacré à ces activités de vérification et de codage peut constituer un avantage important pour les enquêtes. Les progrès réalisés dans le domaine du matériel informatique ont conduit à la mise au point de micro-ordinateurs utilisables sur le terrain. Par ailleurs, grâce aux nouveaux logiciels améliorés spécialement conçus pour l'administration des questionnaires, l'entrevue assistée par ordinateur est maintenant devenue une option envisageable. Les bureaux nationaux des statistiques de certains pays industrialisés ont évalué le recours à cette technique et certains l'utilisent régulièrement (Nicholls et Groves 1986; Lyberg 1985; Denteneer et coll. 1987; Bench et coll. 1994). Ces systèmes présentent l'avantage de réduire les risques d'erreurs en simplifiant le cheminement et en rejetant les entrées inexactes, illogiques ou incohérentes. En outre, ils permettent de stocker les données de très nombreuses entrevues et de les transférer ensuite simplement dans un ordinateur central à la fin de chaque période de travail, éliminant ainsi l'étape de la saisie.

Malgré les avantages apparents de cette technologie, son adoption se butte à des réticences surprenantes dans les pays en développement. Plusieurs raisons peuvent être invoquées pour expliquer ce problème. Premièrement, le coût initial peut paraître prohibitif pour des pays aux ressources limitées. Deuxièmement, il n'y a eu jusqu'à maintenant que peu de tentatives de validation de cette méthode en conditions réelles, et les données quantitatives sur ses avantages et ses inconvénients, qui pourraient en permettre la comparaison avec les méthodes de collecte classiques, sont donc encore limitées (Reitmaier 1985; Ferry et Cantrelle 1988; Forster et coll. 1991). Nous présentons ci-après les résultats d'une étude comparative des deux méthodes de collecte et de traitement des données utilisées au cours d'une étude démographique réalisée sur la côte kényane.

2. ENQUÊTE SUR LA MORTALITÉ DES ADULTES

Cette enquête a été réalisée dans le cadre d'une série d'études démographiques et épidémiologiques portant sur 60,000 habitants de la côte du Kenya. La population étudiée et les méthodes d'enquête démographique utilisées ont été décrites ailleurs (Snow et coll. 1994). En résumé, un recensement initial de la population est suivi d'un contrôle des données de l'état civil réalisé au moyen de visites des ménages individuels effectuées aux six semaines et de recensements semestriels de la population entière. Au cours d'un nouveau recensement de la population effectué en novembre 1993, on a procédé à une estimation de la mortalité des adultes à l'aide de méthodes démographiques indirectes (Timaueus 1991). Toutes les femmes âgées de 25 à

44 ans ont été interviewées à l'aide du questionnaire structuré reproduit à la figure 1. Les questions posées étaient fermées et précodées; le questionnaire comportait des branchements par sauts logiques ainsi qu'une question de vérification de la cohérence.

Vingt-quatre enquêteurs, tous diplômés du niveau secondaire, ont participé à l'étude. Ils étaient tous familiers avec les méthodes d'enquête et de recensement, ayant déjà reçu une formation officielle aux techniques d'enquête sur le terrain et justifiant tous de une à cinq années d'expérience pratique. Ils ont en plus reçu deux jours supplémentaires de formation à l'administration du questionnaire sur la mortalité des adultes. Pendant l'étude, les enquêteurs ont été répartis en deux équipes, chacune placée sous la direction d'un enquêteur principal. Les questionnaires remplis étaient vérifiés à la fin de chaque journée par les chefs d'équipe, puis transmis aux préposés à la saisie. On utilisait un logiciel FoxPro (version 2.0) qui reproduisait à l'écran le questionnaire utilisé. Les données ont été entrées deux fois par deux préposés différents et les dossiers ainsi complétés ont été comparés pour dépister les erreurs de saisie qui ont par la suite été corrigées. Les dossiers ont ensuite été soumis à une vérification logique et à des contrôles de vraisemblance et de cohérence; il s'agissait par exemple de relever les omissions, les codes erronés (p. ex., emploi de lettres autres que "Y" ou "N") et les discordances entre les dates indiquées, l'âge des répondantes et la date de l'enquête (figure 1; questions 5 et 6), et de s'assurer que la somme des réponses aux questions 7, 9, 11 et 13 concordait avec la réponse à la question 15.

3. TEST DE COLLECTE DES DONNÉES ASSISTÉE PAR ORDINATEUR

3.1 Matériel et logiciel

Une version antérieure du logiciel d'enquête avait été élaborée pour le Psion Organiser II (Forster et coll. 1991). Ce modèle comportait un écran aux dimensions limitées à 16 caractères sur deux lignes, mais utilisait un clavier complet. L'ordinateur Psion Series 3 utilisé pour la présente étude offre de nouvelles possibilités: l'écran est beaucoup plus grand (40 caractères sur 8 lignes) et autorise l'affichage graphique. L'appareil reste tout de même petit (165 × 85 × 22 mm) et ne pèse que 265 g, piles comprises (2 piles AA). Les dispositifs de stockage peuvent contenir jusqu'à un mégaoctet. Le clavier QWERTY comporte 58 touches. Le transfert des données entre le Psion Series 3 et un ordinateur personnel classique s'effectue à l'aide d'une simple opération de copie.

Le logiciel a été mis au point à l'aide du langage de programmation intégré du Psion: OPL. Le questionnaire-papier est reproduit sous une forme structurée, dans un fichier texte. La définition du questionnaire comprend un mélange de questions et de commandes (p. ex., pour le branchement par saut et le contrôle de vraisemblance). Les contrôles internes de vraisemblance comprennent ceux mis

Evaluation de l'utilisation des ordinateurs de poche pour la réalisation d'enquêtes démographiques dans les pays en développement

D. FORSTER et R.W. SNOW¹

RÉSUMÉ

Les enquêtes à grande échelle réalisées dans les pays en développement peuvent fournir un instantané extrêmement utile de l'état d'une collectivité, mais les résultats produits reflètent rarement la réalité actuelle puisqu'ils ne sont souvent diffusés que plusieurs mois ou même des années après la collecte des données. Ces retards sont en partie attribuables au temps nécessaire à la saisie, au codage et au traitement des données. Or, il est désormais possible de faire la saisie sur les lieux mêmes de l'enquête, grâce aux ordinateurs de poche. Les erreurs sont également moins nombreuses puisqu'une vérification automatique permet de rejeter les entrées illogiques ou incohérentes dès l'étape de l'administration du questionnaire. Le présent article porte sur l'utilisation d'une telle méthode d'enquête assistée par ordinateur pour la collecte de données démographiques au Kenya. Même si les coûts initiaux de la mise sur pied d'un tel système sont élevés, ses avantages sont évidents: les erreurs qui peuvent se glisser dans les données recueillies, malgré l'expérience du personnel technique, peuvent être réduites à des niveaux négligeables. Dans les situations où la rapidité d'exécution est essentielle, où les employés sont nombreux et où on a recours au précodage des réponses pour la collecte systématique de données sur une période de temps prolongée, l'entrevue assistée par ordinateur pourrait s'avérer économique à long terme, tout en améliorant radicalement la qualité des données à court terme.

MOTS CLÉS: Ordinateurs de poche; enquêtes démographiques; Psion.

1. INTRODUCTION

On procède régulièrement dans les pays en développement à des enquêtes à grande échelle qui peuvent rejoindre des dizaines de milliers de répondants (p. ex., recensements nationaux, enquêtes démographiques et sondages sur l'état de santé des populations). Ces études ont pour objectif de fournir rapidement des informations à jour sur les populations et leur état de santé, aux fins de l'évaluation et de la planification. Leur très grande portée exige la participation d'un grand nombre de formateurs, d'intervieurs, de surveillants, de préposés à la saisie et de gestionnaires. L'Enquête mondiale sur la fécondité (EMF 1986) et les enquêtes nationales sur la démographie et la santé (DHS Kenya 1989) sont des exemples de telles enquêtes fondées sur des questionnaires. La comparaison des dates publiées du début des enquêtes BMF dans 12 pays africains et des dates de publication des premiers rapports nationaux (tableau 1) montre le temps qui peut s'écouler avant que les données ne deviennent accessibles aux planificateurs intéressés (EMF 1986). En moyenne, il faut patienter pendant 45,6 mois avant qu'un rapport final ne soit diffusé. Les retards accusés sont sans doute dus en partie à des problèmes logistiques inhérents aux pays en développement, mais la mécanique du traitement des données a aussi sa part de responsabilités. L'enquête sur la démographie et la santé récemment réalisée au Kenya a employé cinq préposés à la saisie, deux surveillants et un commis au contrôle pour traiter les données de 8,343 entrevues dans les ménages. La collecte des données a commencé en février 1989 et la première ébauche du rapport final a été diffusée sept mois plus tard (DHS Kenya 1989).

Tableau 1
Résumé chronologique des enquêtes mondiales sur la fécondité réalisées dans 12 pays africains
(source: EMF 1986)

Nombre de mois écoulés du début de l'enquête à la publication du rapport	Pays	Nombre d'entrevues	Date du début de l'enquête	Date du premier rapport	Nombre de mois écoulés du début de l'enquête à la publication du rapport
30	Bénin	4,018	12/1981	06/1984	30
63	Cameroun	8,219	01/1978	04/1983	63
52	Ghana	6,125	02/1979	06/1983	52
52	Côte d'Ivoire	6,270	08/1980	12/1984	52
34	Kenya	8,100	08/1977	06/1980	34
52	Lesotho	3,603	08/1977	12/1981	52
41	Mauritanie	3,500	01/1981	06/1984	41
49	Maroc	5,800	04/1980	05/1984	49
35	Nigéria	9,727	10/1981	09/1984	35
38	Sénégal	3,985	05/1978	07/1981	38
40	Soudan (Nord)	3,115	12/1978	04/1982	40
61	Tunisie	4,123	05/1978	06/1983	61

¹ D. Forster, Department of Tropical Medicine, University of Oxford, John Radcliffe Hospital, Headington OX3 9DU, England; R.W. Snow, CRC - Research Unit, Kenyan Medical Research Institute, P.O. Box 230, Kilifi, Kenya.

- ROBINSON, J.G., AHMED, B., DAS GUPTA, P., et al. (1993). Estimation of population coverage in the 1990 United States census based on demographic analysis. *Journal of the American Statistical Association*, 88, 1061-1079.
- ROBINSON, J.G., WORD, D.L., et SPENCER, G. (1991). Uncertainty for models to translate 1990 census concepts into historical racial classifications. 1990 Decennial Census, Preliminary Research and Evaluation Memorandum (PREM) No. 81, Demographic Analysis Evaluation Project D8.
- ROSENWAIKE, I., et LOGUE, B. 1983. Accuracy of death certificate ages for the extreme aged. *Demography*, 20(4), 569-585.
- SHRESTHA, L.B. (1993). Age Misreporting and its Effects on Old-Age Population and Death Registration Estimates: United States, 1970-1990. Unpublished doctoral dissertation, University of Pennsylvania, Population Studies Center.
- SHRYOCK, H.S., et SIEGEL, J.S. (1976). *The Methods and Material of Demography*. Orlando, Florida: Academic Press, Inc. (Harcourt Brace Jovanovich, Publishers): Studies in Population Series.
- SIEGEL, J.S. (1974). Estimates of coverage of the population by sex, race, and age in the 1970 census. *Demography*, 11(1), 1-23.
- SIEGEL, J.S. (1976). New estimates of the number of centenarians in the United States. *Journal of the American Statistical Association*, 71, 559-566.
- SPRAGUE, T.B. (1880-81). Explanation of a new formula for interpolation. *Journal of the Institute of Actuaries*, 22, 270.
- UNITED STATES BUREAU OF THE CENSUS (1972). *General Population Characteristics*. 1970 Census of Population and Housing. Final Report PC(1)-B1. U.S. Summary. Washington, DC: U.S. Government Printing Office.
- UNITED STATES BUREAU OF THE CENSUS (1973). *The Medicare Record Check: an Evaluation of the Coverage of Persons 65 Years of Age and Over in the 1970 Census*. 1970 Census of Population: Evaluation and Research Program, PHC(E)-7. Washington, DC: U.S. Government Printing Office.
- UNITED STATES BUREAU OF THE CENSUS (1974). *Estimates of Coverage of Population by Sex, Race, and Age: Demographic Analysis*. 1970 Census of Population: Evaluation and Research Program, PHC(E)-4. By Jacob S. Siegel. Washington, DC: U.S. Government Printing Office.
- UNITED STATES BUREAU OF THE CENSUS (1975). *Accuracy of Data for Selected Population Characteristics as Measured by the 1970 CPS-Census Match*. 1970 Census of Population: Evaluation and Research Program, PHC(E)-11. Washington, DC: U.S. Government Printing Office.
- UNITED STATES BUREAU OF THE CENSUS (1976). *Demographic Aspects of Aging and the Older Population in the United States*. Current Population Reports. Series P-23, No. 59. By J.S. Siegel. Washington, DC: U.S. Government Printing Office.
- UNITED STATES BUREAU OF THE CENSUS (1983). *General Population Characteristics*. 1980 Census of Population and Housing. Final Report PC80-1-B1. United States Summary. Washington, DC: U.S. Government Printing Office.
- UNITED STATES BUREAU OF THE CENSUS (1984a). *Demographic and Socioeconomic Aspects of Aging in the United States*. Current Population Reports. Series P-23, No. 138. By J.S. Siegel and M. Davidson. Washington, DC: US Government Printing Office.
- UNITED STATES BUREAU OF THE CENSUS (1984b). Census of Population: 1980. Race detail file. 100% count. Table IV: modified counts (OMB-consistent) by age, race, and sex. Unpublished tabulations.
- UNITED STATES BUREAU OF THE CENSUS (1988). *The Coverage of Population in the 1980 Census*. 1980 Census of Population and Housing: Evaluation and Research Reports, PHC80-E4. By R.E. Fay, J.S. Passel and J.G. Robinson. Washington, DC: U.S. Government Printing Office.
- VAUPEL, J.W. (1993). Verbal presentation at Research Workshop on Oldest Old Mortality, Duke University, Durham, North Carolina. March, 1993.
- WILKIN, J.C. (1981). Recent trends in the mortality of the aged. *Transactions of the Society of Actuaries*. Vol. XXXIII, 11-62. WOLFENDEN, H.H. (1954). *Population Statistics and their Compilation*. Revised Edition. The University of Chicago Press: published for the Society of Actuaries.
- WORD, D.L., et SPENCER, G. (1991). Age, sex, race, and Hispanic origin information from the 1990 census: a comparison of census results with results where age and race have been modified. 1990 CHS-L-74. Draft dated August 1991.
- ZELNIK, M. (1969). Age patterns of mortality of American Negroes: 1900-02 to 1959-61. *Journal of the American Statistical Association*, 64, 433-451.

détails sur ces modifications, voir U.S. Bureau of the Census (1984b). Les données des tableaux modifiées sont classées selon la race, le sexe, les années d'âge (0 à 99) et le groupe d'âge (100 ans et plus). Aux fins de notre recherche, nous avons utilisé les groupes raciaux modifiés, compte tenu du nombre considérable de personnes qui ont été transférées du groupe racial résiduel aux catégories blanche ou noire.

1.C Recensement de 1990

La publication des tableaux du recensement de 1990 par le U.S. Bureau of the Census n'est pas encore terminée. Les statistiques déjà publiées présentent toutefois un certain nombre de problèmes qui compliquent leur comparaison avec les recensements antérieurs et les autres sources de données. Trois problèmes sont évidents: classification des 9,3 millions de personnes placées dans une catégorie raciale résiduelle non spécifiée; incohérences dans la déclaration de l'âge et modification des méthodes d'attribution de l'âge pour les personnes qui ont omis de fournir ce renseignement.

Un dossier modifié intitulé MARS (Modified Age and Race Statistics) a été préparé par le Census Bureau pour régler les deux premiers problèmes (Word et Spencer 1991). Les modifications du recensement de 1990 ont porté sur les micro-données. Des méthodes d'imputation spéciales (hot-deck) ont servi à attribuer une race particulière aux personnes qui avaient choisi la catégorie "autre non précisé". Cette redistribution est appliquée aux dossiers individuels des tableaux détaillés entièrement révisés du recensement de 1990 (Robinson, Word et Spencer 1991). Nous avons encore une fois choisi d'utiliser les statistiques modifiées, classées selon la race, le sexe et les années d'âge. Notre décision d'utiliser les statistiques modifiées des recensements de 1980 et de 1990 n'allait pas de soi; pour un examen plus détaillé de la question, voir Shrestha (1993).

2. Le système d'enregistrement des décès

Les statistiques annuelles nationales sur les décès que nous avons utilisées pour notre étude sont venues du National Center for Health Statistics (NCHS). Les données de 1970 à 1988 ont été tirées des bandes de données du NCHS obtenues auprès du ICPHS (NCHS 1970-1988). Elles sont classées selon la race (noire ou blanche), le sexe et les années d'âge (0 à 124 ans; catégorie ouverte pour les 125 ans et plus). Comme les bandes de données de l'année civile 1989 et des trois premiers mois de 1990 n'ont pas encore été diffusées, nous avons élaboré une méthode pour en estimer la distribution. Les statistiques finales sur la mortalité pour 1989 classées selon la race et le sexe ont été publiées par le NCHS en 1992. Les données présentées par groupes d'âge ont été réparties en années d'âge en fonction de la répartition des données de 1988. La répartition selon le mois du décès s'est fondée sur les données des registres de l'état civil (NCHS 1989). Les estimations de la distribution des décès en 1990 sont fondées sur les rapports provisoires mensuels sur la mortalité du NCHS (1990). Ces données provisoires ont été classées en années d'âge en

À cause des erreurs contenues dans les tableaux officiels, nous avons utilisé des tableaux corrigés inédits obtenus auprès du U.S. Bureau of the Census. Ces données statistiques modifiées étaient exemptes des trois types d'erreurs susmentionnés. Les données sont classées selon la race (blanche ou noire), le sexe, les années d'âge (de 0 à 94 ans) et les groupes d'âge (95-99 ans et 100 et plus). Pour répartir les données du groupe des 95-99 ans en années d'âge, nous avons utilisé des distributions de l'âge moyen par sexe et par race des recensements de 1960 et de 1980 pour les blancs, et des recensements de 1950 et 1980 pour les noirs (les données en années d'âge ne sont pas disponibles pour ce groupe d'âge chez les noirs dans le recensement de 1960).

1.B Recensement de 1980

Les tableaux officiels du recensement de 1980 sont présentés dans le document Series B – U.S. Summary of the 1980 Census (U.S. Bureau of the Census 1983). Au cours de ce recensement, un grand nombre de gens (environ 6,8 millions) ont choisi de fournir une réponse de leur cru à la question concernant leur race au lieu de choisir une des catégories préétablies. Comme le recensement de 1980 est le seul à contenir une catégorie raciale résiduelle, il n'est pas directement comparable aux autres sources de données (registres de l'état civil, recensements antérieurs, etc.). Le Census Bureau a produit un dossier modifié conforme aux catégories raciales historiques (U.S. Bureau of the Census 1984b). La méthode d'adaptation des données comportait une redistribution des macro-données sur la race fondée sur un tableau détaillé croisant la race et l'origine hispanique et découlant des données échantillons et des données intégrales du recensement. Nous décrivons en détails ci-après les modifications apportées par le Census Bureau.

Pour les 219,8 millions de personnes qui ont choisi l'une des 14 catégories raciales préétablies, aucun ajustement n'a été fait. Deux catégories de personnes (6,7 millions au total) ont choisi de fournir une description de leur cru: les personnes d'origine hispanique (5,8 millions) et celles d'une autre origine (0,9 million). Des méthodes d'ajustement différentes ont été élaborées pour ces deux groupes. Les personnes d'origine hispanique ont été réparties entre les catégories blanche et noire (et non dans celles des autochtones ni des asiatiques ou des insulaires du Pacifique). Toutes les personnes d'origine mexicaine ont été replacées dans la catégorie blanche. Les personnes d'origine porto-ricaine et cubaine et toutes les autres personnes d'origine hispanique ont été replacées dans des groupes modifiés, blancs ou noirs, dans des proportions identiques au choix des Hispaniques de même origine qui avaient précisé la race blanche ou noire dans leur formulaire. Les calculs ont été effectués par cellules d'âge-sexe-comté. Les personnes d'une autre origine ont été réparties entre trois groupes raciaux modifiés (blancs, noirs et autres) dans des proportions correspondant à celles observées dans les cellules d'âge-sexe-comté de leur État de résidence. Ces proportions étaient fondées sur un échantillon de données tirées du recensement de 1980. Pour de plus amples

principalement par une tendance, chez les personnes de ce groupe, à exagérer leur âge au moment des recensements par rapport aux données inscrites dans les certificats de décès. Il s'ensuit de cette tendance que les taux de mortalité des groupes d'âge supérieurs à 65 ans pour les Afro-américains donneront vraisemblablement lieu à de graves sous-estimations. Il est effectivement possible que les courbes des taux de mortalité des blancs et des noirs se croisent à un âge avancé, mais il serait dangereux de fonder une telle conclusion uniquement sur les données des recensements et des registres de l'état civil. La discordance entre ces deux sources de données est tout simplement trop grande pour nous permettre de faire une estimation fiable des taux de mortalité pour les groupes d'âge avancé.

7. CONCLUSION

Les erreurs de couverture et de contenu relevées dans les systèmes de recensement des personnes et des décès aux Etats-Unis laissent planer un doute sérieux sur la qualité de ces données en ce qui concerne les aînés. Nous avons évalué la concordance des données provenant de ces deux sources pour les populations blanche et afro-américaine, en concentrant notre attention sur les personnes âgées de 60 ans et plus, où les tendances affichées par les taux de mortalité ont la plus grande incidence sur les programmes sociaux et où les données sont les plus problématiques. L'analyse intercensitaire des cohortes laisse constater des discordances liées à l'âge entre les deux sources de données pour les deux périodes: 1970-1980 et 1980-1990.

Afin de déterminer quelles combinaisons de degré de couverture et de tendance à la déclaration d'un âge erroné produirait les résultats empiriques, nous avons réalisé une série de simulations. Les discordances observées dans les données américaines sont examinées à la lumière des résultats de ces simulations et des données de la documentation spécialisée concernant la nature des erreurs de couverture et de contenu dans les sources de données.

Les données concernant les blancs pour la période inter-censitaire 1980-1990 laissent constater une concordance remarquable. Leur qualité, jusqu'à 95 ans, s'approche de celle des données de la Suède et des Pays-Bas, deux pays dotés de registres de la population extrêmement efficaces. Les données portant sur les blancs pour la décennie 1970-1980 laissent voir plus de discordances. L'explication la plus plausible de cette différence est le sous-dénombrement relatif net du recensement de 1970 combiné à la tenue de statistiques plus complètes sur les décès. En conséquence, les estimations de la mortalité chez les aînés qui utilisent des numérateurs tirés des certificats de décès et des dénominateurs tirés du recensement de 1970 risquent de surestimer la mortalité.

Les données concernant les Afro-américains laissent voir une tendance différente. Au-delà de 70 ans, la population recensée diminue graduellement comparativement à la population prévisible tant en 1980 qu'en 1990. Il semble que cette discordance soit principalement due à une tendance à exagérer l'âge lors des recensements, par

Cette recherche a été réalisée au Population Studies Center de la University of Pennsylvania, avec l'aide financière du National Institute of Aging (AG10168) et du Boettner Institute of Financial Gerontology. Nous remercions Irma Elo, Douglas Ewbank, Shiro Horiuchi et J. Gregory Robinson pour leurs commentaires utiles sur notre projet et sur le texte du présent article. Merci également à J. Gregory Robinson et au U.S. Bureau of the Census qui nous ont fournis les données censitaires inédites et les tableaux de données sur les migrations internationales.

REMERCIEMENTS

rapport aux données portées aux certificats de décès. Il s'ensuit de cette tendance que les taux de mortalité des Afro-américains plus âgés risquent d'être sérieusement sous-estimés. Il est possible que les courbes des taux de mortalité des noirs et des blancs se croisent à un âge avancé, mais il serait imprudent de fonder une telle conclusion uniquement sur les données des recensements et des registres de l'état civil.

ANNEXE A

Source: Shrestha (1993)

Trois sources principales de données ont été utilisées dans notre étude: 1) recensements de 1970, 1980 et 1990; 2) registres officiels des décès; 3) statistiques sur l'immigration nette. Nous décrivons ci-après ces sources et les ajustements que nous avons apportés aux données.

1.A Recensement de 1970

Les tableaux officiels des caractéristiques démographiques de base de la population en 1970 sont présentés dans le document Series B - U.S. Summary of the 1970 Census (U.S. Bureau of the Census 1972). Nous savons que ces données officielles contiennent des erreurs importantes qui risquent de biaiser notre enquête sur le recensement des aînés aux Etats-Unis. La première est un surdénombrement évident des personnes centenaires. Alors que 106,000 personnes ont été classées dans cette catégorie ouverte, des analyses démographiques indirectes donnent plutôt à conclure que le nombre correct oscillerait plutôt entre 3,000 et 8,000 (Siegel 1974; Siegel et Passel 1976). Ce surdénombrement semble avoir été causé par une interprétation erronée d'une des questions du formulaire plutôt que par une tendance systématique à se déclarer faussement dans la catégorie des centenaires. Le deuxième problème découle d'une erreur de classification de la population en groupes raciaux dans les tableaux du recensement total qui touche 21,000 personnes âgées de 65 ans et plus. Finalement, le recensement officiel a omis plus de 23,000 personnes (de tous âges) dont les données ont été retrouvées après la publication des tableaux officiels.

Coale et Kisker (1990) préférèrent la première explication. Ils relèvent que les populations reconstituées à partir des données sur les décès à l'aide de méthodes à *r* variables (Preston et Coale 1982) sont trop petites par rapport aux données censitaires de 1980 pour les plus de 65 ans, ce qui porte à conclure que les décès n'ont pas tous été portés aux registres. Ils relèvent également que chez les Afro-américains d'âge avancé, on observe un déficit dans les décès déclarés par rapport aux dossiers Medicare.

Toutefois, ces deux observations pourraient également s'expliquer par une exagération de l'âge lors des recensements (et dans les dossiers Medicare) par rapport aux données des certificats de décès. Une comparaison directe des certificats de décès de 1960 et des données se rapportant aux mêmes sujets dans le recensement de 1960 (NCHS 1968; Hambrigh 1969) tend à confirmer l'existence d'une telle tendance. Pour les hommes comme pour les femmes, le total des décès chez les plus de 50 ans, lorsque ces décès sont classés selon l'âge déclaré au recensement, ne s'écarte pas de plus de 1% du total des décès classés selon l'âge inscrit dans les certificats de décès. Toutefois, pour l'en-semble des personnes de 65 ans et plus, le total des décès classés selon l'âge au recensement dépasse de 15,4% celui des décès classés selon l'âge inscrit dans le certificat de décès dans le cas des femmes, et de 7,1% dans le cas des hommes. Pour les personnes de 75 ans et plus, les écarts atteignent 23,3 et 17,8% respectivement et pour le groupe des 85 ans et plus, ils sont de 39,2 et 17,6% respectivement. Ces écarts importants de l'âge déclaré entre les recensements et les certificats de décès peuvent expliquer la tendance à la baisse des rapports observée dans les graphiques des figures 2A et 2B. Elio et Preston (1994) ont calculé que les valeurs de R_x pour les Afro-américains, entre les périodes 1950-1960 et 1960-1970, périodes qui encadrent la comparaison des données censitaires et celles des certificats de 1960. Ils démontrent que si les âges au décès sont 'corrigés' pour correspondre à ceux déclarés lors des recensements, la tendance à la baisse des rapports est éliminée.

Les raisons qui poussent les Afro-américains à déclarer un âge plus avancé lors des recensements, comparativement à l'âge inscrit dans le certificat de décès, ne sont pas claires. Cette tendance n'apparaît pas avant le recensement de 1940, c'est-à-dire le premier réalisé après l'adoption de la Loi sur la sécurité sociale. Ce recensement laisse constater un large excédent d'Afro-américains appartenant aux groupes d'âge 65-69 ans et 70-74 ans, et un déficit des personnes du groupe d'âge 50-64 ans (Elio et Preston 1994). Comme le rappelle Wolfenden (1954:56), 'Les distorsions dans les données sur les noirs étaient si marquées qu'il a fallu procéder à une nouvelle répartition préliminaire de ces populations (et de ces décès) entre 55 et 69 ans, aux fins de la préparation des tables de survie.' Cet excédent persiste, même s'il s'atténue graduellement, jusque dans les recensements les plus récents (comme le montre la figure 2). Quelle qu'en soit la raison, nous croyons que l'explication principale des écarts importants observés entre les données censitaires et les données des certificats de décès pour la population noire américaine s'expliquent

le programme postcensitaire au cours duquel des données individuelles du recensement sont comparées à des données provenant d'autres systèmes.

6.1.2 Période intercensitaire 1980-1990

Comme mentionné précédemment, l'évolution des rapports dans le cas des blancs (en particulier les femmes) laisse constater une très grande concordance des données, de loin meilleure que dans la plupart des pays européens. Notre étude semble confirmer l'hypothèse de Vaupel (1993) selon laquelle la population blanche des États-Unis présenterait des taux de mortalité plus bas, au-delà de l'âge de 80 ans, que dans tous les autres pays industrialisés. Toutefois, la prudence est de mise. Si nos méthodes laissent voir une grande concordance des données entre les recensements et les certificats de décès pour la période 1980-1990, rien n'indique que ces données soient nécessairement exactes. Condran et coll. (1991) décrit une situation dans laquelle une tendance à déclarer des âges erronés donne une série de rapports de 1,00 pour tous les âges. Par ailleurs, la méthode intercensitaire ne permet pas de dépister les sujets qui font systématiquement des déclarations erronées de leur âge. Comme le soulignait Horiuchi (1993), une personne qui exagère une première fois son âge – pour être admise à une école ou être autorisée à travailler plus jeune, pour éviter le service militaire lorsqu'elle approche de l'âge limite supérieur ou pour bénéficier plus tôt de la sécurité sociale, du programme Medicare ou d'une pension – pourra avoir tendance à continuer d'exagérer son âge par la suite. Une telle possibilité ne peut être ignorée, même si nous sommes incapables de la mesurer directement.

6.2 Résultats pour les noirs

L'évolution des rapports en fonction de l'âge est beaucoup plus régulière dans le cas des Afro-américains (voir les figures 2A et 2B). Ils commencent à diminuer vers 70 ans pour les deux sexes dans les deux périodes et pour- suivent cette chute jusqu'aux âges avancés (jusqu'à 100 ans en 1970-1980, les rapports se maintiennent typiquement au-dessus de 1,00 en 1970-1980, et légèrement au-dessus de 1,00 pour les hommes en 1980-1990. L'écart observé entre les rapports pour des groupes d'âge donnés entre les résultats de 1970-1980 et de 1980-1990 n'est pas étranger au sous-dénombrement relatif qui a caractérisé le recensement de 1970. Comme nous l'avons déjà mentionné, il est vraisemblable qu'un tel sous-dénombrement se soit produit également dans le cas des blancs. Ce sous-dénombrement ne suffit pas toutefois à expliquer la tendance à la baisse persistante des rapports au-delà de 70 ans pour les deux périodes. Deux raisons principales peuvent permettre d'expliquer ce déclin des séries de rapports pour les Afro-américains:

- 1) L'inscription des décès aux registres d'état civil n'est pas faite avec la même efficacité que les recensements;
- 2) L'exagération de l'âge est plus importante dans les recensements que dans les certificats de décès.

Figure 1A. Rapports intercensitaires de la population observée sur la population prévue: blancs, 1970-1980.

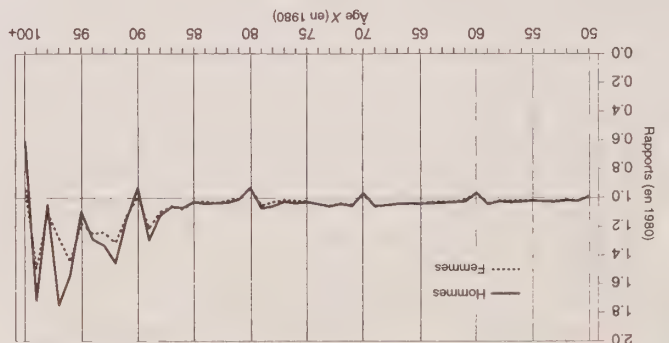


Figure 1B. Rapports intercensitaires de la population observée sur la population prévue: blancs, 1980-1990.

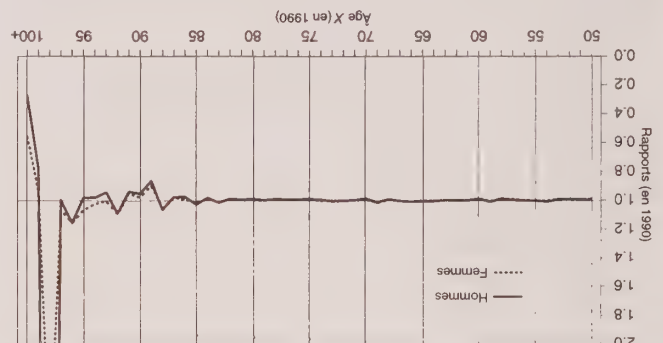


Figure 2A. Rapports intercensitaires de la population observée sur la population prévue: noirs, 1970-1980.

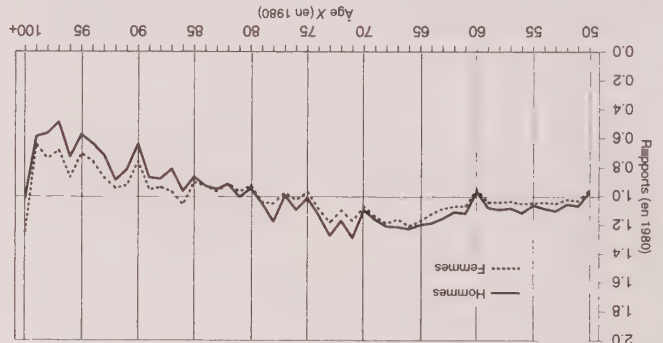
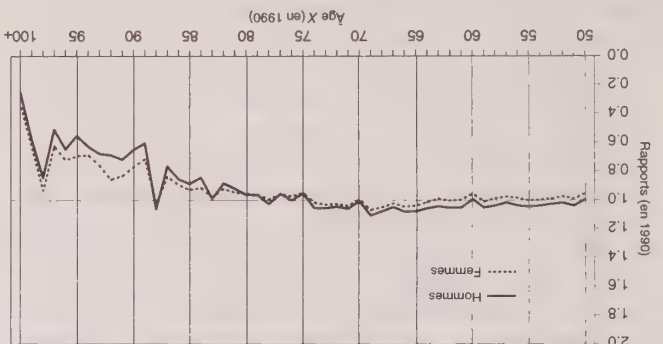


Figure 2B. Rapports intercensitaires de la population observée sur la population prévue: noirs, 1980-1990.



Pour toutes les combinaisons de races et de périodes, l'effet des tendances relatives à l'âge sur les rapports est pratiquement le même pour les hommes et pour les femmes. Dans tous les cas, le degré de discordance augmente avec l'âge, même si on ne commence à observer un écart systématique et significatif par rapport à 1,00 qu'à partir de 95 ans dans le cas des blancs en 1980-1990. On observe une discontinuité nette dans beaucoup de ces séries à l'âge de 100 ans, ce qui reflète le caractère idiosyncratique de la déclaration de l'âge et les méthodes d'ajustement utilisées par le Census Bureau dans le cas des centenaires.

6.1 Résultats pour les blancs

6.1.1 Période intercensitaire 1970-1980

Comme l'indique la figure 1A, les rapports de la population blanche ont tendance à se maintenir au-dessus de 1,00 au cours de cette période et ils augmentent avec l'âge (jusqu'à 100 ans). Cette configuration pourrait être le résultat de divers types d'erreurs de données dont les deux plus plausibles sont:

- 1) un sous-dénombrement dans le recensement de 1970 par rapport au recensement de 1980 et au recensement des décès;
- 2) des probabilités à peu près égales d'exagération de l'âge dans les certificats de décès et lors des deux recensements.

Nous croyons que la première explication risque davantage d'être la bonne. Si l'évolution des rapports en fonction du vieillissement découlait de tendances semblables à l'exagération de l'âge déclaré dans les certificats de décès et les recensements, on pourrait s'attendre à une situation comparable au cours de la décennie 1980-1990, notamment du fait que le recensement de 1980 figure dans les deux comparaisons. On ne s'attendrait pas non plus à voir les proportions culturelles à déclarer un âge erroné disparaître soudainement. Pourtant, le graphique de l'évolution des rapports pour les blancs au cours de la décennie 1980-1990 (figure 1B) laisse constater une remarquable concordance, bien meilleure que dans la plupart des pays européens et équivalant à celle observée en Suède et dans les Pays-Bas, deux pays dotés de registres de la population extrêmement efficaces (Condran et coll. 1991). La concordance des données observées pour la décennie 1980-1990 est également beaucoup plus grande que celle obtenue dans d'autres pays de langue anglaise: Angleterre et pays de Galles, Canada, Australie et Nouvelle-Zélande.

Notre préférence pour la première explication est également motivée par le fait que le Census Bureau a conclu que le recensement de 1980 était plus complet que celui de 1970 (U.S. Bureau of the Census 1988; Robinson et coll. 1993). Cette conclusion est fondée en partie sur l'analyse démographique et n'est donc pas entièrement indépendante du genre de preuve que nous examinons. Toutefois, l'analyse démographique du Bureau fait grand usage de groupes d'âge plus jeunes que ceux auxquels nous intéressons ici. En outre, la conclusion selon laquelle la couverture des recensements a été améliorée est également confirmée par

indiqué pourra également subir l'influence: 1) des erreurs de couverture dans l'un ou l'autre ou les deux recensements; 2) des erreurs (en plus ou en moins) dans les registres des décès ou les statistiques de l'immigration; 3) des déclarations erronées (concernant l'âge, la race, etc.) dans l'une ou l'autre ou l'ensemble des sources de données (Ewbank 1981; Shryock et Siegel 1976; Condran et coll. 1991). Le rapport des populations observées sur les populations prévues est un outil de diagnostic utile lorsqu'il est possible d'attribuer les écarts par rapport à 1.00 à l'effet de ces erreurs sous-jacentes. Toutefois, il ne constitue pas un outil très précis puisque différents types d'erreurs peuvent produire le même genre d'écarts. Il peut néanmoins permettre un choix plus facile entre diverses solutions concurrentes.

5. EFFETS DES ERREURS SUR LES RAPPORTS DES POPULATIONS OBSERVÉES SUR LES POPULATIONS PRÉVUES

Certains types d'erreurs ont un effet direct visible sur la formule du rapport lui-même (effets qui ont été confirmés par les simulations que nous avons réalisées). Pour simplifier la démonstration, désignons par R_x dans l'équation (2) le rapport de la population observée sur la population prévue pour l'âge x et plus au deuxième recensement. On peut distinguer les possibilités principales suivantes d'erreurs de couverture et leurs répercussions sur la configuration des rapports en fonction de l'âge:

- 1) Si $N_x(1)$ et D sont également complets et que $N_x(2)$ est assorti d'un degré de complétude relative de $C(2)$, alors la configuration des rapports en fonction de l'âge sera constante et son niveau sera $C(2)$.
- 2) Si $N_x(2)$ et D sont également complets et que $N_x(1)$ est assorti d'un degré de complétude relative de $C(1)$, alors la configuration des rapports en fonction de l'âge sera:
 - a) supérieure à 1.00 et augmentera en fonction de l'âge si $C(1) < 1.00$;
 - b) inférieure à 1.00 et diminuera en fonction de l'âge si $C(1) > 1.00$.

La raison pour laquelle ces erreurs induisent un effet relatif à l'âge sur le rapport R_x est qu'une erreur proportionnelle donnée dans $N_x(1)$ engendre des erreurs proportionnelles de plus en plus grandes dans le dénominateur (l'un positif et l'autre négatif) tendent l'un vers l'autre en valeur absolue. Cette égalisation est due au fait qu'une proportion plus grande des membres de chaque cohorte meurent pendant la période intercensitaire à mesure que l'âge avance.

- 3) Si $N_x(1)$ et $N_x(2)$ sont également complets et que $C(D)$, alors la configuration des rapports en fonction de l'âge sera:

- a) supérieure à 1.00 et augmentera en fonction de l'âge si $C(D) > 1.00$ (c.-à-d., si les certificats de décès sont plus complets que les dénombrements des deux recensements);
- b) inférieure à 1.00 et diminuera en fonction de l'âge si $C(D) < 1.00$.

Ici encore, on observe une tendance relative à l'âge puisque une erreur proportionnelle égale dans D engendrera des erreurs proportionnelles plus grandes dans le dénominateur à mesure que les deux éléments de ce dernier tendent l'un vers l'autre en termes absolus.

On peut mieux comprendre les effets des déclarations erronées de l'âge en examinant les éléments qui composent cette formule. Shrestha (1993) et Condran et coll. (1991) ont introduit diverses erreurs dans des ensembles simulés, dépourvus d'erreurs, typiques des conditions démographiques actuelles existant aux États-Unis et aux Pays-Bas respectivement. Ils ont démontré qu'une tendance nette à exagérer l'âge qui se limiterait à deux recensements produirait des rapports oscillant aux alentours de 1.00 jusqu'à un âge avancé, pour chuter ensuite à des valeurs très faibles. La diminution du rapport à des valeurs inférieures à 1.00 s'explique ici encore par le fait qu'une erreur dans un des éléments du dénominateur (dans ce cas, l'inflation de $N_x(1)$ par une exagération de l'âge) engendre des effets disproportionnés dans le dénominateur. Même si la chute rapide de la distribution par âge peut entraîner une inflation plus grande de $N_x(2)$ par rapport à $N_x(1)$, l'inflation du dénominateur finira tôt ou tard par excéder celle du numérateur et les rapports diminueront. (À titre d'exemple, voir la figure 1 de Condran et coll. 1991).

Un problème d'exagération de l'âge limité aux recensements des décès donnera un rapport supérieur à 1.00 et qui augmentera avec l'âge. Dans ce cas, le dénominateur est trop petit (sa composante négative est trop grande) et le déficit proportionnel augmente avec l'âge. L'existence d'une tendance identique à exagérer l'âge dans les certificats de décès et les recensements donnera également des rapports qui finiront par augmenter avec l'âge. Ce résultat important est robuste dans la mesure de l'erreur introduite (Condran et coll. 1991). Il découle du fait que les distributions par âge chutent de plus en plus rapidement à mesure que l'âge avance, de sorte qu'un pourcentage identique de personnes qui exagèrent leur âge véritable entraînera un pourcentage plus grand d'erreurs dans les distributions par âge aux âges très avancés. Autrement dit, $N_x(2)$ est assorti d'un facteur d'inflation plus grand que $N_x(1)$. Dans ce cas, une certaine inflation de $N_x(1)$ verra ses effets sur le dénominateur compensés par l'inflation de D .

6. RÉSULTATS

L'analyse intercensitaire des cohortes a porté sur quatre groupes de sexe-race des États-Unis, au cours des périodes 1970-1980 et 1980-1990. Les figures 1 et 2 illustrent les rapports calculés de la population observée sur la population prévue pour des groupes d'âge précis en fonction de la race, du sexe et de la période intercensitaire.

Une étude au cours de laquelle on a comparé un échantillon de certificats de décès datant de mai à août 1960 aux données du recensement de 1960 constitue la meilleure évaluation de la concordance des données sur l'âge déclaré contenues dans les tableaux de recensements et dans les registres de l'état civil – sans doute une des sources les plus importantes d'erreurs de contenu influant sur notre test de cohérence (NCHS 1968; Hambrigh 1969). Même si les données ont été recueillies avant la période visée par notre projet, les résultats de cette étude mettent en lumière un problème qui risque d'exister encore aujourd'hui. Les auteurs de l'étude ont constaté ce qui suit: 1) dans le cas des blancs, les données des deux sources concordent assez bien même lorsque l'âge augmente, mais pour les personnes de couleur, la concordance est moins nette; 2) lorsqu'il y a discordance entre les deux sources de données, la différence d'âge dans le cas des blancs est généralement inférieure à un an, mais cette différence est typiquement supérieure à un an dans le cas des personnes de couleur, en particulier pour celles âgées de 45 ans et plus; 3) pour les blancs de tous âges et pour les personnes de couleur âgées de moins de 45 ans, l'âge indiqué sur le certificat de décès est typiquement plus avancé que celui déclaré au moment du recensement. Par contre, pour les personnes de couleur âgées de 45 ans et plus, l'âge indiqué sur le certificat de décès est en moyenne moins avancé que celui déclaré lors du recensement.

Les auteurs de cette étude n'ont pas été en mesure de déterminer laquelle des deux sources donnait l'âge "réel", Rosenwaike et Logue (1983) ont cherché à vérifier les données sur l'âge indiqué dans les certificats de décès des personnes âgées de 85 ans et plus pour la période de 1968 à 1972. Ils ont à cette fin tiré un échantillon de certificats de personnes décédées à un âge très avancé en Pennsylvanie et aux New Jersey. Ils ont ensuite comparé ces données à celles du recensement manuscrit de l'année 1900. Au total, 1,429 comparaisons ont été établies: 960 concernant des personnes de race blanche et 496 concernant des personnes de couleur.

Ils ont constaté que la concordance des renseignements sur l'âge entre les deux sources diminuait à mesure que l'âge augmentait pour les deux groupes raciaux. Des différences frappantes ont été observées entre les groupes raciaux. La concordance était élevée dans le cas des blancs, sauf pour les personnes âgées de 100 ans et plus. Toutefois, pour les personnes de couleur, la concordance était sensiblement moins bonne. Les auteurs ont par ailleurs relevé qu'à l'intérieur de chaque groupe racial, le sexe avait peu d'incidence sur la concordance des données.

4. MÉTHODE INTERCENSITAIRE D'ÉVALUATION DE LA QUALITÉ DES STATISTIQUES SUR LES AÎNÉS

La présente analyse permet de déterminer l'ampleur de la discordance des sources de données sur les aînés à l'aide d'une méthode intercensitaire fondée sur les cohortes. La taille prévisible d'une cohorte d'âge ouverte dans le second

recensement peut être estimée à partir de sa taille dans le premier recensement et des décès survenus au sein de cette cohorte au cours de la période intercensitaire, après l'ajustement tenant compte des migrations (Condran, Himes et Preston 1991). Le recours à une catégorie ouverte permet l'observation de la tendance du rapport tout en atténuant l'effet des valeurs extrêmes dues à des erreurs dans des groupes d'âge particuliers. Cette méthode est insensible aux erreurs commises dans la déclaration de l'âge des personnes décédées ou des personnes recensées dont l'âge est supérieur à celui qui marque le début de la catégorie ouverte.

En utilisant les données censitaires et celles portant sur les décès et les migrations pour une période intercensitaire donnée, l'analyse intercensitaire des cohortes nous permet d'estimer la taille prévisible de chaque cohorte d'âge ouverte du recensement suivant. Les statistiques susmentionnées, classées par années d'âge, par sexe et par race (blanche et noire), ont servi à calculer l'équation suivante pour la population prévue à l'époque du second recensement:

$$N_x(2) = N_{x-10}(1) - D_{x-10}(1) + M_{x-10}(1) \quad (1)$$

où

$$N_x(2) = \text{la population prévue d'âge } x \text{ et plus, lors du second recensement, réalisé 10 ans après le premier;}$$

$$N_{x-10}(1) = \text{la population recensée d'âge } x - 10 \text{ et plus au temps 1, soit au premier recensement;}$$

$$D_{x-10}(1) = \text{les décès survenus pendant la période intercensitaire dans la cohorte des personnes d'âge } x - 10 \text{ et plus (au moment du premier recensement);}$$

$$M_{x-10}(1) = \text{l'immigration légale nette intercensitaire dans la cohorte des personnes d'âge } x - 10 \text{ et plus (au moment du premier recensement).}$$

On peut également calculer à l'aide d'une méthode analogue la population prévue d'un âge donné (par opposition à celle d'âge x et plus). Dans l'un ou l'autre de ces cas, il est ensuite possible de calculer le rapport de la population observée, dénombrée lors du recensement suivant, sur la population prévue (après simplification et en présupposant que la migration nette est nulle) à l'aide de la formule suivante:

$$R_x = \frac{N_x(2)}{N_{x-10}(1) - D_{x-10}(1)} \quad (2)$$

Le changement de taille de la cohorte mesuré entre deux recensements successifs ne peut être dû qu'aux décès et aux migrations. Un rapport de 1,00 indiquerait donc une concordance complète entre les sources de données. (Noter toutefois qu'un tel rapport n'est pas nécessairement un signe d'exactitude des données. Ainsi, si l'âge d'une personne donnée a systématiquement été augmenté de n années, la méthode ne permettra pas de relever cette déclaration erronée). Dans les faits, toutefois, le dénombrement

diminuées des nombres estimés de décès dans la cohorte et des migrations, aux données censitaires (voir à ce propos le résumé dans Robinson et coll. 1993, et Himes et Clogg 1992). Les méthodes statistiques comparent un groupe de personnes tiré d'une source de données de rechange (p. ex., le Current Population Survey) aux données individuelles du recensement. Une troisième méthode consiste à comparer les données censitaires concernant les aînés aux dénombrements des personnes inscrites dans les dossiers du programme fédéral d'assurance-maladie (Medicare).

Les évaluations du recensement de 1970 réalisées par le U.S. Bureau of the Census (1973, 1974, 1975) ont conduit à un certain nombre de conclusions générales concernant la population des aînés. Premièrement, l'erreur nette (combinaison des erreurs de couverture et de contenu) est plus importante dans le cas des aînés que des personnes plus jeunes. Deuxièmement, les taux d'erreur nets ont été plus élevés dans le cas des femmes que dans celui des hommes, par suite du plus grand nombre d'erreurs commises dans la déclaration de l'âge. Cependant, les taux bruts d'omission (qui ne constituent qu'un élément de l'erreur nette) étaient plus élevés dans le cas des hommes. Troisièmement, les taux d'erreur nets, d'omission bruts et d'erreur dans la déclaration des caractéristiques démographiques sont beaucoup plus élevés dans le cas des américains de race noire que de ceux de race blanche. Quatrièmement, il semble que les statistiques officielles contiennent une très grande quantité d'erreurs dans la déclaration de l'âge. Par exemple, il est intéressant de noter que pour les quatre groupes de race-sexe des personnes âgées de 65 à 69 ans en 1970, les estimations tirées de l'analyse démographique laissent conclure à une surestimation nette dans les données censitaires tandis que l'étude fondée sur les données du Medicare laisse transparaître d'importantes omissions dans les données censitaires (de 2,1 % dans le cas des femmes blanches à 12,6 % dans le cas des hommes noirs et de ceux des autres catégories raciales). Cette comparaison donne à conclure qu'en plus des taux d'omission bruts dans le dénombrement des personnes âgées de 65 à 69 ans, d'autres erreurs encore plus grandes (surtout, semble-t-il, une exagération de l'âge chez les personnes de moins de 65 ans) agissent dans le sens contraire pour augmenter les estimations du dénombrement net des personnes appartenant à ces groupes d'âge. Il découle notamment de cette situation que les caractéristiques d'une portion substantielle de la population classée dans le groupe des 65 ans et plus lors du recensement se rapportent en fait à des personnes âgées de moins de 65 ans (U.S. Bureau of the Census 1976).

Comparativement aux données censitaires de 1970, les taux d'erreurs nets observés en 1980 dans la plupart des groupes d'âge-race-sexe ont été sensiblement moins élevés. Toutefois, comme le faisait observer le Bureau of the Census (1988), les résultats du programme postcensitaire (Post Enumeration Program, ou PEP) et de la Housing Unit Enumeration Duplication study de 1980 donnent à conclure qu'une part considérable du total des dénombrements censitaires, soit plus de 1,1 % selon toute vraisemblance, contient des personnes déjà recensées. Les données

laissent supposer que le taux de duplication était beaucoup moindre au cours des recensements antérieurs. Ainsi, les améliorations relevées dans la couverture nette du recensement de 1980 semble malheureusement en partie attribuables à un problème de duplication des données (U.S. Bureau of the Census 1988:10).

Le Census Bureau compte procéder à des évaluations complètes de la qualité du recensement de 1990, mais la diffusion des résultats obtenus est demeurée jusqu'à maintenant fragmentaire. Il semble que le problème de sous-dénombrement des données brutes ait été moins grave en 1980 qu'en 1990 (Robinson et coll. 1993), mais cette différence pourrait être due à un taux de duplication plus élevé dans le recensement de 1980. On peut formuler un certain nombre de généralisations concernant la tendance au sous-dénombrement net dans le recensement de 1990, dans le cas des aînés. Premièrement, confirmant la tendance historique, les estimations de l'erreur nette pour les Afro-américains dépassent largement celles des blancs. La différence la plus importante s'observe dans le cas des hommes âgés de 60 à 64 ans. Le taux de sous-dénombrement net pour les hommes de race noire atteint 10,3 %, alors qu'il n'est que de 2,6 % chez les hommes de race blanche (différence de 7,7 points). Deuxièmement, alors qu'on observe un sous-dénombrement pour tous les groupes d'âge chez les hommes, il existe des problèmes de surdénombrement dans plusieurs des groupes de femmes. Finalement, comme le soulignent Robinson et coll. (*ibid*), les tendances de la couverture nette sont généralement comparables dans les trois derniers recensements pour chaque groupe de race-sexe.

Les statistiques officielles sur les décès produites par le National Center for Health Statistics constituent la source de base de données sur la mortalité annuelle aux États-Unis. Ces chiffres sont généralement utilisés sans ajustement pour tenir compte du sous-dénombrement et des erreurs de déclaration qui se glissent dans les certificats de décès. Toutefois, on présume en général que le système de recensement des décès est pratiquement complet (Wilkin 1981; U.S. Bureau of the Census 1984a; National Center for Health Statistics 1968) même si aucun test national n'a été fait pour confirmer ce diagnostic depuis la mise en place de la Death Registration Area en 1933. Cette supposition s'appuie sur les exigences légales strictes en vigueur concernant le recensement des décès ainsi que sur la nécessité, pour les survivants, de prouver le décès aux fins des arrangements funéraires, du règlement des droits successoraux et de la collecte des prestations d'assurance (U.S. Bureau of the Census 1984a; Wilkin 1981). Les calculs réalisés par Coale et Kisker (1990) donnent toutefois à penser qu'il existe un problème de sous-dénombrement des décès, en particulier chez les aînés. Dans le cas des personnes de couleur, par exemple, le total des décès portés aux registres de l'état civil est inférieur de 7 % à celui des hommes âgés de plus de 80 ans en 1980, et de 10 % dans le cas des femmes. Toutefois, il est possible que l'écart relevé soit dû à des différences dans l'âge déclaré entre les deux sources plutôt qu'à un problème de sous-dénombrement.

méthode d'analyse intercensitaire s'étale du 1^{er} avril au 31 mars, alors que les registres des décès utilisent l'année civile. En outre, les registres des décès et les recensements américains utilisent l'âge au dernier anniversaire au lieu de l'année de la naissance. Nous avons corrigé les deux problèmes en répartissant les décès sur des graphes triangulés du temps et de l'âge correspondant aux "années de recensement" débutant le 1^{er} avril. Par exemple, les décès déclarés dans l'intervalle 1970 et celle du 1^{er} avril 1971 et concernant des personnes âgées de 60 ans (à leur dernier anniversaire) au moment du recensement peuvent être répartis en quatre catégories: 1) personnes de 60 ans décédées au cours de l'année civile 1970; 2) personnes de 60 ans décédées au cours de l'année civile 1971; 3) personnes de 61 ans décédées au cours de l'année civile 1970; 4) personnes de 61 ans décédées au cours de l'année civile 1971. À l'aide des données sur les dates des décès tirées des bandes du NCHS, nous avons répartis les décès sur des triangles du temps et de l'âge correspondant à l'année de recensement débutant le 1^{er} avril 1970. Ce faisant, nous avons présumé que les décès compris dans chaque triangle étaient répartis uniformément. Cette supposition est rendue nécessaire par l'absence de données fiables sur la naissance pour la plupart des cohortes visées par notre étude, données grâce auxquelles il nous aurait été possible de répartir les décès plus précisément entre les cohortes de naissance adjacentes. Pour une description plus détaillée de cette méthode, voir Shrestha (1993).

2.4 Statistiques sur l'immigration nette

Nous avons utilisé des statistiques inédites sur l'immigration nette obtenues auprès du U.S. Bureau of the Census. Si la qualité des statistiques américaines sur l'immigration est généralement mise en doute (Hill 1985), l'estimation de la taille de la population des aînés est passablement indépendante des variations dans les estimations de la migration intercensitaire. Cette robustesse découle à la fois du nombre moindre de migrants nets chez les personnes plus âgées, comparativement aux plus jeunes, et du rôle plus important de la mortalité dans la diminution du nombre de personnes âgées, comparativement à la migration nette. Par exemple, les données sur l'immigration nette font état d'un apport de 64 hommes de race noire pour la cohorte des personnes âgées de 75 ans et plus (en 1970) pendant la décennie de 1970 à 1980. À titre de comparaison, plus de 141,000 décès ont été enregistrés dans cette cohorte au cours de la même période.

Les estimations concernant les migrations des résidents sans document sont exclues des ensembles de données sur l'immigration nette, mais elles seront prises en compte dans l'interprétation des résultats. Cette exclusion a été motivée par un certain nombre de facteurs. Les estimations de la taille et de la répartition selon l'âge et le sexe des immigrants illégaux (illégal aliens) varient largement par suite de l'insuffisance d'instruments de collecte de données aux États-Unis. Cependant, même les estimations les plus exagérées du nombre de migrants sans document

sont négligeables par rapport aux décès en ce qui concerne les aînés.

Nous avons décrit un certain nombre d'ajustements apportés aux données de base: utilisation de compilations de données censitaires inédites de 1970 à cause d'un sur-dénombrément marqué de la population des centenaires dans les statistiques officielles; utilisation de compilations modifiées pour la race des recensements de 1980 et de 1990 et exclusion des estimations concernant les populations d'étrangers sans document. Pour juger des effets de ces ajustements sur nos résultats, nous avons procédé à plusieurs analyses de sensibilité utilisant des données non corrigées. Même si nous n'avons observé que de légères différences entre les résultats obtenus avec les données officielles et les données corrigées (sauf pour les 100 ans et plus), les analyses intercensitaires de cohortes utilisant des données non corrigées ont généralement laissé voir des écarts plus grands dans les résultats définitifs, confirmant dans l'ensemble la pertinence des corrections apportées.

3. SOURCES D'ERREURS DANS LES RECENSEMENTS ET LES REGISTRES DE DÉCÈS

Les erreurs dans les données démographiques ont été classées en deux catégories: erreurs de couverture et erreurs de contenu. La couverture s'entend de la mesure dans laquelle les personnes ou les événements faisant partie d'un ensemble défini dans un système de données particulières sont pris en compte. Le contenu s'entend de la qualité de l'information effectivement compilée sur les personnes ou les événements. L'un ou l'autre de ces deux types d'erreurs peut créer des discordances entre l'évolution intercensitaire de la taille d'une cohorte et les décès intercensitaires. Toutefois, si les deux sources de données sont marquées par le même taux net d'omissions, la concordance entre les deux sera maintenue. En outre, dans ces conditions, les taux de mortalité resteront également exacts.

Même si les cas d'erreurs dans la déclaration de l'âge suivent la même tendance dans les recensements et dans les certificats de décès, cela ne permettra pas, en règle générale, de maintenir la concordance des changements de taille des cohortes entre les deux sources de données. En effet, comme les taux de mortalité augmentent avec l'âge, la distribution par âge des décès chez les aînés est "plus vieille" que la distribution par âge de la population. Par exemple, si on place par erreur 10% des personnes et des décès du groupe d'âge 75-79 dans le groupe d'âge 80-84, l'incidence proportionnelle de cette erreur sur les données démographiques sera plus grande que l'incidence proportionnelle sur les taux de mortalité. Cette tendance à déclarer un âge erroné aurait pour effet de baisser les taux de mortalité et influencerait également sur la cohérence des tests que nous utilisons.

Le Census Bureau a utilisé des méthodes démographiques et statistiques pour estimer la qualité de la couverture des recensements. Les méthodes démographiques comparent les estimations du nombre réel de naissances,

par années d'âge. Nous sommes maintenant en mesure de combler cette grave lacune puisque nous avons traité les bandes de données pour chacun des décès enregistrés aux États-Unis de 1970 à 1988. (Les chiffres pour 1989 (année complète) et pour 1990 (janvier à mars seulement) ont été estimés à partir des données par groupes d'âge publiées par le National Center for Health Statistics et des données de l'année 1988. Voir l'annexe A pour plus de détails.) Ces bandes sont produites par le National Center for Health Statistics (NCHS) et sont diffusées par le Inter-University Consortium for Political and Social Research (ICPSR). Pour les années que nous avons examinées, elles portent sur un nombre approximatif de 50 millions de décès.

2. POPULATIONS ET DONNÉES

2.1 Contexte général

Trois sources principales de données ont été utilisées: 1) recensements nationaux du U.S. Bureau of the Census pour les années 1970, 1980 et 1990; 2) données annuelles sur les décès compilées par le NCHS; 3) estimations inédites sur l'immigration nette obtenues auprès du U.S. Bureau of the Census. La description détaillée de nos sources de données est jointe au présent article dont elle constitue l'annexe A. Il nous paraît cependant utile de décrire ci-après brièvement ces données et les ajustements importants que nous leur avons apportés.

2.2 Recensements

Nous utilisons des données censitaires classées en fonction de la race (noire/blanche) et du sexe, et par années d'âge (avec une classe ouverte pour les 100 ans et plus). Les compilations englobent l'ensemble des résidents des 50 états et du district de Columbia; elles comprennent les personnes placées en établissements, les Américains en déplacement temporaire à l'étranger et les étrangers dont la résidence habituelle (légale ou non) est située aux États-Unis (sauf les militaires et le personnel diplomatique). Sont par contre exclus les Américains qui résident à l'étranger pour une période prolongée et les ressortissants étrangers en visite temporaire aux États-Unis. Les statistiques officielles ne sont pas ajustées pour tenir compte du sous-dénombrement (p. ex., résidents légaux et immigrants sans document non recensés).

Le choix de l'expression "population résidente" signifie que les tableaux de recensement comprennent à la fois les résidents légaux et les immigrants sans document. Sur l'ensemble des personnes sans document qui résidaient aux États-Unis à l'époque du recensement de 1970, il semble qu'une proportion négligeable ait été dénombrée. Ainsi, la population des résidents légaux était approximativement comparable à la population résidente totale lors du recensement de 1970. Par contre, lors du recensement de 1980, le U.S. Bureau of the Census a estimé que, pour la toute première fois, un nombre significatif de personnes sans document avaient été dénombrées. Cette population a été évaluée à 2,06 millions de personnes. De ce nombre,

2.3 Registres des décès

Les registres américains des décès portent sur l'ensemble des décès recensés dans les 50 états et dans le district de Columbia, classés selon la race et le sexe et par années d'âge (jusqu'à 125 +). Pour permettre une comparaison avec les données des recensements, les décès des non résidents (ressortissants étrangers et citoyens américains résidant à l'étranger) ont été exclus. Aucune correction n'est faite pour tenir compte du sous-dénombrement des décès ou des décès dont la description sur le certificat comporte des erreurs. Nous avons relevé deux problèmes qui influent sur l'utilisation de notre

population blanche.

On reconnaît que les tableaux officiels du recensement de 1970 contiennent des erreurs parmi lesquelles la plus évidente est le surdénombrement marqué du nombre de personnes âgées de 100 ans ou plus. Même si le recensement a dénombré 106,000 personnes appartenant à ce groupe d'âge, les estimations démographiques indirectes laisseraient plutôt conclure à un nombre oscillant entre 3,000 et 8,000 personnes, le chiffre estimatif le plus plausible étant fixé à 4,800 (Siegel 1974; Siegel et Passel 1976; U.S. Bureau of the Census 1974). Nous utilisons les données inédites du U.S. Bureau of the Census pour le recensement de 1970, où cette surestimation des centenaires a été corrigée. Cette utilisation des estimations corrigées est motivée par deux raisons: premièrement, à défaut d'une correction, l'excédent serait suffisamment important pour biaiser les résultats pour les groupes plus âgés. Deuxièmement, il semble que cette surestimation n'ait pas été due à des déclarations erronées de l'âge par les intéressés, mais plutôt à un malentendu qui a porté certaines personnes à confondre les colonnes destinées à l'inscription du mois et de l'année de la naissance (Siegel et Passel 1976). Au cours des recensements de 1980 et de 1990, un grand nombre de personnes ont choisi de fournir une réponse de leur cru à la question concernant la race, au lieu de choisir une des catégories préétablies. Sur la population totale, 6,8 millions de personnes, en majeure partie d'origine hispanique, ont choisi cette option en 1980, et 9,3 millions on fait de même en 1990. Les tableaux officiels de ces recensements ne sont donc pas directement comparables à d'autres sources de données puisqu'ils sont seuls à prévoir une catégorie raciale résiduelle. Pour autoriser une comparaison avec d'autres sources de données, le Census Bureau a modifié les données censitaires de 1980 et de 1990 pour qu'elles soient conformes aux catégories historiques de groupes raciaux. Le Bureau a en outre procédé, en 1990, à une "correction" du problème lié à l'âge (pour plus de détails, voir Word et Spencer 1991). Nous avons décidé d'utiliser les statistiques du Bureau de 1980 et de 1990 modifiées pour la race aux fins de notre étude. Ce choix a été motivé par le nombre considérable de personnes qui auraient été exclues autrement, en particulier dans la

Concordance des données censitaires et des registres de l'état civil concernant les aînés aux États-Unis: 1970-1990

LAURA B. SHRESTHA et SAMUEL H. PRESTON¹

RÉSUMÉ

Les lacunes et les erreurs relevées dans les données censitaires et les registres de l'état civil aux États-Unis laissent planer un doute sérieux sur la qualité des données concernant les aînés et le recensement des décès dans ce pays. Dans le présent article, nous évaluons la concordance des données provenant de ces deux sources pour les populations blanche et afro-américaine. Nous étudions en particulier les aînés (60 ans et plus), où les tendances de la mortalité ont la plus grande incidence sur les programmes sociaux et où les problèmes de données sont les plus graves. Au moyen de l'analyse intercensitaire des cohortes, nous déterminons les discordances en fonction de l'âge entre les deux sources pour deux périodes: 1970-1980 et 1980-1990. Les anomalies relevées quant à la nature de la couverture et aux erreurs de contenu sont analysées à la lumière des données de la documentation spécialisée. Les données sur la population afro-américaine montrent de très nombreuses discordances pour la période 1970-1990, qui sont vraisemblablement attribuables à une exagération de l'âge déclaré lors des recensements, par rapport aux données des registres de l'état civil. On relève également des anomalies pour la population blanche dans la période intercensitaire 1970-1980. Selon nous, ces anomalies sont principalement dues à un sous-dénombrement dans le recensement de 1970, par rapport à celui de 1980 et au recensement des décès. Par contre, les données de 1980-1990 concernant les blancs, et en particulier les femmes, sont très cohérentes, beaucoup plus même que dans la plupart des pays européens.

MOTS CLÉS: Erreurs de déclaration au sujet de l'âge; couverture; mortalité; évaluation des recensements; recensement des décès; qualité des données; croisement des courbes du taux de mortalité, États-Unis.

1. INTRODUCTION

Les méthodes classiques d'estimation de l'ampleur de la mortalité dans les pays développés utilisent des données provenant de deux sources distinctes. Les données sur les décès tirées des registres de l'état civil servent habituellement de numérateurs des taux de mortalité. Les dénominateurs correspondent pour leur part habituellement au dénombrement des personnes vivantes réalisé lors des recensements. L'exacitude des taux ainsi calculés dépend donc de la qualité des données provenant de ces deux sources. Nous présentons ci-après les résultats d'un test de la qualité des données américaines correspondant à deux périodes intercensitaires: 1970-1980 et 1980-1990. Nous étudions en particulier la concordance des changements relevés dans la taille d'une cohorte d'un recensement à l'autre et du nombre de décès recensés pendant cette période, en procédant aux ajustements voulus pour tenir compte des migrations. Toutes les données sont exprimées en années d'âge, et des tests séparés sont réalisés pour les populations blanche et noire. Nous nous intéressons essentiellement aux aînés (60 ans et plus), où les tendances de la mortalité ont la plus grande incidence sur les programmes sociaux (Preston 1993) et où les problèmes de données sont les plus graves. La population blanche des États-Unis semble artifier des taux de mortalité plus bas, chez les plus de 80 ans, que dans tous les autres pays industrialisés (Vaupel 1993). Si elle s'avérait valide, cette comparaison pourrait avoir une grande incidence

sur l'évaluation de la qualité relative des services de santé. Cependant, la population afro-américaine présente des taux de mortalité encore plus bas que ceux de la population blanche chez les plus de 80 ans, reflétant ainsi le phénomène bien connu du croisement des courbes du taux de mortalité des deux races qui s'observe entre 75 et 85 ans. La fiabilité de ces deux ensembles de données sur la mortalité dépend bien sûr de la qualité des données. Or, on a déjà émis des doutes très sérieux sur la qualité des données concernant la population de race noire (voir à ce propos Zelnik 1969; Coale et Kisker 1990), même si la plupart des observateurs semblent admettre d'emblée la validité du croisement des tendances (Manton et coll. 1986; McCord et Freeman 1990). Dans le cadre de leurs travaux de détermination de nouveaux profils modèles du taux de mortalité dans les pays à faible mortalité, Condran, Himes et Preston (1991) ont fait état de tests semblables de la qualité des données pour 68 périodes intercensitaires dans 18 pays industrialisés. En général, la concordance des données s'est avérée très bonne pour les cohortes de 65 ans au deuxième recensement (66 des 68 ensembles de données ont passé le test de cohérence). La concordance a cependant diminué avec l'âge: la moitié seulement des ensembles de données montraient une concordance à 85 ans et la proportion des données concordantes passait à 15% à 95 ans (Condran et coll. 1991: tableau 7). Les États-Unis ne faisaient pas partie des pays qui ont fait l'objet de ces tests parce que leurs données publiées sur la mortalité ne sont pas classées

¹ Laura B. Shrestha, The World Bank, Human Development Department, 1818 H Street, N.W., Washington, DC 20433, U.S.A.; Samuel H. Preston, Population Studies Center, University of Pennsylvania, 3718 Locust Walk, Philadelphia, PA 19104, U.S.A.

- comparaisons, plus il y a de chances que certaines d'entre elles soient déclarées à tort statistiquement significatives. Dans ce cas, nous employons d'autres mesures statistiques pour contrôler le taux global d'erreur du processus décisionnel.
- Nous avons procédé à l'analyse de manière à ce que les constatations sur la "famille" complète des 28 combinaisons ou sur les paramètres de régression logistique maintiennent l'intervalle de confiance de 90% (une norme du Census Bureau) dans tous les cas. L'intervalle de confiance de 90% des comparaisons par paire a été ajusté au moyen de la méthode C de Dunnett permettant la comparaison par paire contrastante des estimations du panel (Hockberg et Tamhane 1987). La méthode d'inférence simultanée de Bonferroni a servi à évaluer la signification statistique des paramètres de régression logistique.
- ## BIBLIOGRAPHIE
- ARMSTRONG, J.S., et LUSKE, E.J. (1987). Return postage in mail surveys: A meta analysis. *Public Opinion Quarterly*, 51 (1) 233-248.
- DILLMAN, D.A., CLARK, J., et SINCLAIR, M. (1993). Effects of questionnaire length, respondent-friendly design, and a difficult question on response rates for occupant-addressed census mail surveys. *Public Opinion Quarterly*.
- DILLMAN, D.A., SINCLAIR, M., et CLARK, J. (1992). Mail-back response rates for simplified decennial census questionnaires. *Proceeding of the Section on Survey Research Methods, American Statistical Association*, 776-783.
- DILLMAN, D.A. (1991). The design and administration of mail surveys. *Annual Review of Sociology*, 17, 225-249.
- DILLMAN, D.A. (1978). *Mail and Telephone Surveys: The Total Design Method*. New York: Wiley-Interscience.
- DUNCAN, W.J. (1979). Mail questionnaires in survey research: A review of response inducement techniques. *Journal of Management*, 5, 39-55.
- FOX, R.J., CRASK, M.R., et KIM, J. (1988). Mail survey response rate: A meta-analysis of selected techniques for inducing response. *Public Opinion Quarterly*, 52, 467-491.
- HARVEY, L. (1987). Factors affecting response rates to mailed questionnaires: A comprehensive literature review. *Journal of the Market Research Society*, 29, 3, 342-353.
- HEBERLEIN, T., et BAUMGARTNER, R. (1978). Factors affecting response rates to mailed questionnaires: A quantitative analysis of the published literature. *American Sociological Review*, 43, 447-462.
- HOCHBERG, Y., et TAMHANE, A.C. (1987). *Multiple Comparison Procedures*. New York: John Wiley and Sons.
- KANUK, L., et BERENSON, C. (1975). Mail surveys and response rates: A literature review. *Journal of Marketing Research*, 12, 440-453.
- KULKA, R.A., HOLT, N.A., CARTER, W., et DOWD, K.L. (1991). Self reports of time pressures, concerns for privacy and participation in the 1990 Mail Census. *Proceedings of the 1991 Annual Research Conference*. U.S. Bureau of the Census, 33-54.
- LINSKY, A.S. (1975). Stimulating responses to mailed questionnaires: A review. *Public Opinion Quarterly*, 39, 82-101.
- MISKURA, S.M. (1992). Estimating the Full Cycle Costs for the Simplified Questionnaire Test (SQT), 2KS Memorandum Series, Design 2000, Book I, Chapter 30, #6.
- SCOTT, C. (1961). Research in mail surveys. *Journal of Royal Statistical Society*, 143-205.
- THOMPSON, J.H. (1993). Final Results of the Mail Response Evaluation for the Implementation Test (IT), DSSD 2000 Census Memorandum Series, #E-32.
- WOLTER, Kirk (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

Bien que l'effet individuel (2,5% globalement) de l'enveloppe-réponse affranchie soit légèrement inférieur au taux nécessaire pour présenter une valeur significative, son ordre de grandeur concorde avec la valeur jugée significative dans des recherches antérieures (Armstrong et Luske 1987; Dillman 1978 et 1991). Face aux nombreux résultats de recherche qui ont démontré son utilité, cette technique ne devrait probablement pas être complètement écartée pour raison d'inefficacité. Il semble aussi que l'on puisse établir un lien d'une autre nature entre, d'une part, l'enveloppe affranchie et, d'autre part, la lettre de préavis et le rappel. Lorsque l'enveloppe affranchie est utilisée seule avec la lettre de préavis, son effet est significatif (3,4 points), mais à l'évidence il ne l'est pas (1,6 point) lorsqu'un rappel est inclus dans l'envoi. Le rappel compense l'absence d'enveloppe affranchie, alors que le préavis semble accroître l'effet de l'enveloppe lorsqu'elle y est. C'est peut-être qu'un préavis signalé aux gens d'être à l'affût de la trousse postale sur le recensement et que, une fois celle-ci ouverte, la présence de l'enveloppe-réponse affranchie incite les gens à répondre. Ce lien différentiel avec les envois qui précèdent et ceux qui suivent ne semble pas avoir été examiné dans les recherches antérieures. Il en découle qu'en cas de réception d'une lettre de préavis sans rappel ultérieur, la présence d'une enveloppe-réponse affranchie pourrait augmenter le taux de réponse de manière statistiquement significative, mais revêtir une importance moindre lorsque l'expéditeur utilise aussi une carte de rappel, comme c'était le cas lors du dernier recensement.

Il existe au moins deux obstacles importants à l'application directe des résultats de notre recherche au recensement de l'an 2000. D'abord, il faut reconnaître que les tests actuels sont effectués hors d'une période de recensement. Par le passé, le Census Bureau a obtenu des taux de réponse beaucoup plus faibles dans les années de non-recensement que dans les années de recensement. Par exemple, le National Content Test de 1986 a obtenu seulement un taux de réponse de 49,2% avec un questionnaire de remplacement, tandis que le Recensement de 1990, sans questionnaire de remplacement, a obtenu un taux de réponse de 65%. Un écart aussi prononcé tiendrait à un climat dit "favorable au recensement", explication succincte de l'association de divers éléments, comme l'attention médiatique, la publicité et le sens culturel de la participation qui semblent faire surface chaque décennie au cours de l'année du recensement.

Les taux de réponse obtenus dans nos tests au moyen des cinq composantes qui augmentent le taux de réponse sont beaucoup plus élevés que ceux qu'on obtient généralement en période de non-recensement; ils sont néanmoins à peu près égaux ou peut-être mieux légèrement inférieurs aux résultats obtenus au cours du dernier recensement décennal, où aucune de ces composantes n'a été utilisée. Nous ignorons si l'existence d'un "climat favorable au recensement" peut remplacer les effets de ces composantes ou améliorer le taux de réponse susceptible d'être atteint l'année d'un recensement. Nous ne pensons certainement pas qu'une augmentation de 30 points du taux de réponse au Recensement de l'an 2000 soit possible, car cela signifierait qu'à

Les résultats de l'analyse découlent de deux méthodes distinctes: les comparaisons multiples réalisées entre les taux de réponse postale selon le type d'intervention, et la régression logistique. Chaque méthode a ses avantages pour ce qui est de l'interprétation des données dans un cas, et de l'inférence statistique dans l'autre; c'est pourquoi nous avons utilisé une approche combinée, recourant aux deux.

Dans notre étude, nous avons calculé le taux de réponse national estimé d'un panel donné en divisant le total pondéré du nombre de questionnaires retournés par le total pondéré des formulaires de recensement postés, moins le total pondéré des envois retournés par le maître de poste (généralement des unités vacantes).

Nous avons examiné les résultats des comparaisons multiples des taux de réponse relatifs aux huit types d'intervention afin de déterminer l'importance de l'augmentation du taux de réponse pour chacun d'eux. Ces comparaisons comportaient l'évaluation par paire de tous les types d'intervention, les uns avec les autres ainsi qu'avec le groupe de contrôle.

La méthode de régression logistique est un moyen rapide et efficace de déterminer si l'augmentation du taux de réponse attribuable à chaque composante (mais surtout à leur interaction) résulte d'une variation de l'échantillon-nage ou est bien réelle, et si l'augmentation observée est influencée par la présence d'autres variables. Toutefois, les paramètres estimés ne peuvent être facilement assimilés aux taux de réponse postale. Pour plus de précisions, voir la méthode de régression logistique dans Thompson 1993.

Les taux de réponse ont été calculés pour chaque type d'intervention à l'intérieur de chacune des strates et à l'échelle nationale (strate 1 et strate 2 réunies). Les erreurs-types relatives aux estimations nationales ont été obtenues au moyen de la méthode jackknife du calcul de la variance pour un échantillonnage statistique VPLX. Les erreurs-types pour les estimations relatives aux strates ont été obtenues avec la méthode jackknife de calcul de la variance pour un échantillonnage aléatoire simple.

L'analyse principale a fait appel à la comparaison par paire des écarts observés entre les taux de réponse obtenus pour huit types d'intervention, tant à l'échelle nationale qu'à l'échelle des strates (RTRF et RTRF).

En raison de la diversité des hypothèses testées, l'analyse porte sur la comparaison de toutes les combinaisons possibles (28 au total) entre les huit types d'intervention. Au moyen de la régression logistique, nous avons testé huit paramètres de modèle ou plus afin de déterminer si les résultats étaient statistiquement significatifs. Plus il y a de

ANNEXE
Méthodes d'estimation

peu près 100% du public-cible a répondu. Par conséquent, il reste passablement d'incertitude quant aux répercussions exactes des résultats actuels sur le Recensement de l'an 2000.

3. RÉSULTATS

Deux méthodes analytiques servent à présenter les principales constatations de cette étude. La première fait appel à des comparaisons multiples, par paire, des moyennes des types d'intervention; la deuxième recourt à la régression logistique. Les méthodes d'estimation sont précisées en annexe. Les deux méthodes donnent des résultats cohérents. Les taux de réponse globaux et les erreurs-types globales pour chacun des types d'intervention à l'échelle nationale et des strates sont présentés au tableau 1. Ils varient entre 50,0% pour le groupe de contrôle et 64,3% lorsque les trois principaux effets sont combinés.

3.1 Comparaisons multiples des taux de réponse par la poste

Le tableau 2 présente 28 comparaisons correspondant à tous les appariements possibles des huit types d'intervention. Compte tenu du peu d'espace qu'offre le tableau, nous avons utilisé les abréviations suivantes: C = contrôle, L = lettre de préavis, E = Enveloppe-réponse affranchie, R = Carte de rappel.

Les trois premières comparaisons figurant au tableau 2 indiquent une amélioration du taux de réponse attribuable à chacune des trois principales composantes s'ajoutant au groupe de contrôle, prises séparément. L'amélioration estimée du taux de réponse attribuable à la lettre de préavis s'établissait à 4,2% dans la strate des RTRF, à 6,7% dans la strate des RTRF, et à 6,4% à l'échelle nationale. L'amélioration estimée attribuable à la carte de rappel était de 5,7% dans la strate des RTRF, de 8,3% dans la strate des RTRF, et de 8,0% à l'échelle nationale. Toutes ces améliorations sont significatives. Par conséquent, la principale constatation de cette étude est que tant la lettre de préavis que la carte de rappel ont accru le taux de réponse à l'échelle nationale et des strates. On n'a enregistré aucune amélioration significative avec l'enveloppe-réponse affranchie, ni à l'échelle nationale ni dans les strates.

Nous avons évalué un modèle comportant une composante pour les strates, la lettre de préavis, l'enveloppe-réponse affranchie et la carte de rappel, y compris tous les termes de l'interaction. Il y a également eu modélisation à l'échelle des strates, avec pour seuls paramètres les effets des composantes et leurs interactions.

Les résultats de l'analyse du modèle global indiquent que seuls les principaux effets de la lettre et de la carte de rappel, ainsi que l'élément stratification et les coordonnées à l'origine sont statistiquement significatifs. Compte tenu de ces résultats, une autre modélisation à l'échelle nationale a été effectuée au moyen d'un modèle réduit comportant seulement les principaux effets de la stratification, des composantes individuelles et des interactions entre ces dernières. Les résultats sont présentés au tableau 3 ci-après.

3.2 Analyse de régression logistique

La lettre de préavis, l'enveloppe-réponse affranchie et la carte de rappel ont chacune amélioré le taux de réponse de 6,4, 2,5 et 8,0 points respectivement. L'accroissement de 2,5 points n'est pas statistiquement significatif. On a également établi que les effets des composantes étaient essentiellement additifs ou indépendants les uns des autres. Par rapport au groupe de contrôle, la combinaison lettre de préavis-enveloppe affranchie a amélioré le taux de réponse de 9,8%, tandis que la combinaison enveloppe affranchie-rappel a accru le taux de réponse de 9,5%, et la combinaison lettre-rappel l'a accru de 12,7%. Ensemble, les trois composantes ont amélioré le taux de réponse de 14,3%. Chaque utilisation de la lettre et du rappel a amélioré considérablement le taux de réponse; quant à l'enveloppe-réponse affranchie, elle n'a eu d'effet significatif que lorsqu'elle était utilisée avec une lettre de préavis, sans rappel. La conclusion la plus importante à tirer de cette expérience est que la lettre de préavis et la carte de rappel sont essentielles à l'atteinte d'un taux de réponse élevé et qu'aucune de ces composantes n'élimine l'effet de l'autre.

4. DISCUSSION ET CONCLUSIONS

Les deux modélisations montrent, tant pour le modèle national que pour le modèle des strates, une amélioration significative du taux de réponse résultant de la lettre et du rappel, mais non de l'enveloppe-réponse affranchie. Ces résultats correspondent à ceux qui ont été obtenus pour les comparaisons multiples présentées ci-dessus. Aucune interaction n'était statistiquement significative, signe que les effets des composantes sont essentiellement additifs.

Un i.c. marqué d'un astérisque (*) indique que l'écart est statistiquement significatif à $\alpha = .10$.

Paramètres du modèle	Estimation	I.C. de 90%
Coordonnées à l'origine, β_0	-.61	-.686 à -.545*
Stratification, β_1	.738	.689 à .789*
Lettre de préavis, β_2	.227	.130 à .324*
Enveloppe-réponse affranchie, β_3	.090	-.006 à .186
Carte de rappel, β_4	.291	.194 à .387*
Lettre de préavis - enveloppe-réponse affranchie, β_5	.036	-.101 à .173
Lettre de préavis - carte de rappel, β_6	-.054	-.192 à .083
Carte de rappel - enveloppe-réponse affranchie, β_7	-.043	-.179 à .093
Lettre de préavis - carte de rappel - enveloppe-réponse affranchie, β_8	-.003	-.197 à .191

Tableau 3

Analyse de régression logistique des moindres carrés pondérés, modèle interaction réduit, sans strate composante

Tableau 1
Taux finals – Test de mise en oeuvre – Estimations à l'échelle nationale et des strates

Type d'intervention	Taux de réponse (%) – Estimations et erreurs-types (%)			
	Échelle nationale		Régions à taux de réponse élevé	
	Estimation	Erreur-type	Estimation	Erreur-type
1. Groupe de contrôle	50.0	0.8	51.9	0.9
2. Lettre de préavis seulement	56.4	0.8	58.6	0.9
3. Enveloppe-réponse affranchie seulement	52.6	0.8	54.5	0.9
4. Carte de rappel seulement	58.0	0.8	60.2	0.9
5. Lettre et enveloppe-réponse affranchie	59.8	0.8	62.1	0.9
6. Enveloppe-réponse affranchie et carte de rappel	59.5	0.8	61.8	0.9
7. Lettre de préavis et carte de rappel	62.7	0.8	65.0	0.9
8. Lettre de préavis, enveloppe-réponse affranchie et carte de rappel	64.3	0.8	66.5	0.9

Tableau 2
Écarts dans les taux de réponse – Chaque composante en présence d'une autre composante

Écarts dans les taux de réponse (%) et intervalles de confiance (i.c.) de 90%

Comparaisons expérimentales	Échelle nationale		Régions à taux de réponse faible (RTRF) – 1990		Régions à taux de réponse élevé (RTRF) – 1990	
	Écart	I.C. de 90%	Écart	I.C. de 90%	Écart	I.C. de 90%

1. L – C	6.4	3.3 à 9.5*	4.2	0.9 à 7.5*	6.7	3.2 à 10.2*
2. S – C	2.5	–0.5 à 5.6	1.7	–1.7 à 5.0	2.7	–0.8 à 6.1
3. R – C	8.0	4.9 à 11.1*	5.7	2.4 à 9.1*	8.3	4.9 à 11.7*
4. LS – C	9.8	6.7 à 12.9*	6.8	3.4 à 10.1*	10.2	6.7 à 13.7*
5. SR – C	9.5	6.4 à 12.5*	6.4	3.0 à 9.7*	9.9	6.5 à 13.3*
6. LR – C	12.7	9.6 à 15.7*	9.2	5.8 à 12.5*	13.2	9.7 à 16.6*
7. LSR – C	14.2	11.2 à 17.2*	11.5	8.2 à 14.8*	14.6	11.3 à 18.0*
8. L – S	3.8	0.8 à 6.9*	2.5	–0.9 à 5.9	4.1	0.6 à 7.5*
9. R – L	1.6	–1.5 à 4.8	1.5	–1.9 à 5.0	1.6	–1.96 à 5.10
10. R – S	5.5	2.4 à 8.5*	4.1	0.7 à 7.5*	5.6	2.2 à 9.0*
11. LS – L	3.4	0.3 à 6.5*	2.6	–0.9 à 6.0	3.5	0.03 à 7.0*
12. SR – L	3.1	0.03 à 6.2*	2.2	–1.3 à 5.6	3.2	–0.3 à 6.6
13. LR – L	6.3	3.2 à 9.3*	5.0	1.5 à 8.4*	6.4	3.0 à 9.9*
14. LS – S	7.3	4.2 à 10.3*	5.1	1.7 à 8.5*	7.6	4.1 à 11.0*
15. SR – S	6.9	3.8 à 10.1*	4.7	1.2 à 8.2*	7.2	3.8 à 10.7*
16. LR – S	10.1	7.1 à 13.2*	7.5	4.1 à 11.0*	10.5	7.0 à 13.9*
17. LS – R	1.8	–1.3 à 4.9	1.1	–2.4 à 4.5	1.9	–1.6 à 5.4
18. SR – R	1.5	–1.6 à 4.5	0.7	–2.8 à 4.1	1.6	–1.8 à 5.0
19. LR – R	4.7	1.6 à 7.7*	3.5	–0.02 à 6.9	4.9	1.5 à 8.3*
20. LSR – L	7.9	4.8 à 10.9*	7.3	3.9 à 10.7*	7.9	4.5 à 11.4*
21. LSR – S	11.7	8.7 à 14.7*	9.8	6.4 à 13.3*	12.0	8.6 à 15.4*
22. LSR – R	6.2	3.2 à 9.3*	5.8	2.3 à 9.3*	6.3	2.9 à 9.7*
23. LSR – LS	4.4	1.4 à 7.5*	4.7	1.2 à 8.2*	4.4	1.0 à 7.8*
24. LSR – SR	4.8	1.7 à 7.8*	5.1	1.7 à 8.6*	4.7	1.3 à 8.2*
25. LSR – LR	1.6	–1.4 à 4.5	2.3	–1.1 à 5.8	1.5	–1.8 à 4.8
26. SR – LS	–0.3	–3.3 à 2.7	–0.4	–3.8 à 3.1	–0.3	–3.7 à 3.1
27. LR – LS	2.9	–0.2 à 6.0	2.4	–1.1 à 5.9	2.9	–0.6 à 6.4
28. LR – SR	3.2	0.2 à 6.2*	2.8	–0.6 à 6.2	3.3	–0.1 à 6.6

Un intervalle de confiance marqué d'un astérisque (*) indique que l'écart est statistiquement significatif à $\alpha = 0.10$ (9 chances sur 10 que l'i.c. comprenne l'écart actuel).

Quatrième, le libellé du préavis, "D'ici quelques jours, vous devriez recevoir..." et du rappel, "Vous auriez dû recevoir ces jours derniers..." avait pour objectif d'encourager les destinataires à être à l'affût du formulaire de recensement. Cinquièmement, l'utilisation du papier à en-tête du directeur du Census Bureau et de la carte postale blanche affichant le sceau du Department of Commerce au-dessus du message de rappel avait pour but d'informer le destinataire que le questionnaire provenait du gouvernement et non d'un quelconque groupe tentant de se faire passer pour un organisme gouvernemental, comme cela se fait parfois au moyen d'une mention du genre "Nous vous informons officiellement que..."

L'influence favorable – à supposer qu'il y en ait eu une – de l'enveloppe-réponse affranchie sur le taux de réponse pourrait être attribuable au fait qu'on réussit à persuader le destinataire de la légitimité et de l'importance de la demande qui lui est faite. (Autrement, pourquoi l'expéditeur "gaspillerait-il" un timbre qu'on peut facilement retirer de l'enveloppe et utiliser ailleurs? Le destinataire hésiterait peut-être à jeter un objet qui a de la valeur (p. ex., un timbre non oblitéré)). Le préavis et, dans une certaine mesure, le rappel, peuvent accroître l'effet du timbre en incitant le destinataire à ouvrir l'enveloppe contenant le formulaire de recensement. De plus, le fait de s'apercevoir, une fois qu'on a ouvert l'enveloppe, qu'elle renferme un timbre non utilisé peut inciter à en conserver le contenu, ce qui renforce l'effet du rappel.

Pour que le préavis, l'enveloppe-réponse affranchie et le rappel s'appuient réciproquement, on a jugé important d'utiliser le courrier de première classe. Si l'on avait eu recours à des envois en nombre au tarif réduit et si ces envois avaient été étroitement espacés, il est probable que, dans certains cas, ces derniers ne seraient pas parvenus à destination dans le bon ordre.

En somme, notre test comportait plus que la simple juxtaposition de trois composantes distinctes dont il est fait mention dans la documentation. Ces composantes ont été opérationnalisées de façon à accroître la probabilité que chacune renforce l'effet des autres et à ce qu'elles s'appliquent aux envois postaux à grande échelle. En réalité, nous espérons apprendre si l'une ou plus d'une de ces composantes pouvaient être éliminées sans que le taux de réponse diminue exagérément, ce qui nous montrerait comment réduire le coût des envois liés au recensement.

2. PLAN EXPERIMENTAL

Un plan factoriel, formé des huit combinaisons possibles des trois principaux effets, a été utilisé pour l'expérience. Les types d'intervention ont été les suivants:

- 1) Néant (groupe de contrôle),
- 2) lettre de préavis seulement,
- 3) enveloppe-réponse affranchie seulement,
- 4) carte de rappel seulement,
- 5) lettre et enveloppe-réponse affranchie,

L'univers d'échantillonnage correspondait à l'ensemble des unités de logement situées dans les régions visées par les questionnaires postaux et établies à partir du fichier d'adresses du Census Bureau. Les 449 régions correspondantes aux bureaux de district (BD) du Recensement de 1990 ont été assimilées aux unités géographiques devant servir à définir les strates aux fins du test. Deux strates ont été définies. Compte tenu de la forte corrélation entre le taux relatif aux minorités (on entend par "minorités" toutes les classifications relatives aux populations de race noire et d'origine hispanique) et le taux de réponse par la poste au Recensement de 1990, on a pu réaliser la stratification en classant les BD selon le pourcentage de membres des minorités que l'on y trouvait. Les BD dont une forte proportion de la population se composait de minorités (noire ou d'origine hispanique) et dont le taux de réponse au Recensement de 1990 était faible ont été définis comme des "régions à taux de réponse faible" (RTRF) et formaient la première strate. Les BD restants ont été classés comme des "régions à taux de réponse élevée" (RTRBE) et constituaient la seconde strate.

2.1 Plan d'échantillonnage

Un échantillon de 50,000 unités de logement a été sélectionné, puis divisé en deux strates de 25,000 unités chacune. On a suréchantillonné la strate des RTRF afin d'étudier simultanément divers facteurs liés au sous-dénombrement différentiel, thème que nous n'aborderons pas dans le présent article. Chaque strate a été divisée en huit panels de même taille afin de mettre à l'essai les huit types d'intervention. Un échantillon systématique de 3,125 unités de logement a été sélectionné dans chaque combinaison panel-strate. Une fois qu'une unité de logement était sélectionnée, les sept unités suivantes l'étaient également. Les ménages de chacun des huit groupes ont été assignés au hasard à un type d'intervention différent. L'échantillon a été regroupé, ce qui a permis de réduire la variance d'échantillonnage lors de la comparaison des panels.

La taille de l'échantillon sélectionné pour cette étude a été définie grâce à une vaste opération de simulation de données, laquelle a indiqué que l'échantillon de 50,000 unités était suffisant pour qu'on détecte un écart d'un minimum de 3% dans l'ensemble des comparaisons par paire des types d'intervention.

- 6) enveloppe-réponse affranchie et rappel,
- 7) lettre et rappel, et
- 8) lettre, enveloppe-réponse affranchie et rappel.

1.2 Conception et intégration des éléments d'intervention

La trousse postale renfermant le formulaire de recensement présentait certaines caractéristiques nous permettant de penser que les destinataires pourraient ne pas l'avoir repérée ou ne pas en avoir tenu compte. Par nécessité, cette trousse est seulement adressée à des ménages anonymes; on ne peut utiliser de nom. Or, pour assurer le traitement approprié des questionnaires retournés, il faut que l'adresse exacte du ménage figure sur le questionnaire proprement dit. Dans un recensement de grande envergure, cependant, l'adressage distinct d'une enveloppe extérieure, d'une lettre et d'un questionnaire, et la procédure visant à s'assurer que les bons objets sont insérés dans la bonne enveloppe, posent de sérieux problèmes en matière de contrôle de la qualité. C'est pourquoi il importe d'imprimer l'adresse sur un seul des objets devant être rassemblés en une trousse postale. Nous utilisons donc pour la livraison de la trousse postale une enveloppe à fenêtre laissant voir l'adresse imprimée sur le questionnaire. L'impossibilité de recourir au nom des répondants ainsi que la taille et l'apparence de l'enveloppe à fenêtre donnent malheureusement l'impression que celle-ci ne contient rien d'important ou renferme peut-être même de la publicité-rebut. De plus, des recherches sur les cas de non-réponse au Recensement de 1990 ont révélé que certaines personnes ne se rappelaient pas avoir reçu un questionnaire de recensement par la poste, ou l'avaient vu, mais n'avaient pas ouvert l'enveloppe, deux phénomènes qui peuvent être attribuables à l'apparence de publipostage (Kulka et coll. 1991). Dans ce test, la lettre de préavis et la carte de rappel avaient pour but d'attirer l'attention sur l'enveloppe renfermant le formulaire de recensement. L'objectif a été atteint de cinq manières. Premièrement, on a élaboré le préavis sous forme de lettre et le rappel sous forme de carte postale. On a pensé que les gens seraient plus susceptibles d'examiner deux objets de correspondance s'ils paraissaient différents l'un de l'autre. On a choisi la lettre en guise de préavis, préférant conserver la carte postale pour le rappel, en raison de la commodité de sa présentation.

Deuxièmement, le préavis consistait en une lettre du directeur du Censu Bureau, avec la mention "Aux résidents du", suivie de l'adresse imprimée à l'emplacement habituel sur le papier à lettre. Nous voulions faire savoir aux destinataires que le questionnaire du recensement qui leur arriverait sous peu était spécialement destiné au ménage résidant à cette adresse. L'adresse était également visible par la fenêtre de l'enveloppe, ce qui réglait l'éventuel problème de contrôle de la qualité qu'aurait soulevé l'envoi par la poste du formulaire de recensement regrou-pant des objets de correspondance adressés séparément. Troisièmement, le préavis était censé être livré quelques jours avant l'enveloppe contenant le formulaire de recensement proprement dit, et le rappel devait suivre à peine quelques jours plus tard. Les dates fixées pour ces envois étaient les 21, 24 et 29 septembre respectivement. On avait pensé que, pour être efficace, un rappel (sans question-naire de remplacement) devait arriver dans les quelques jours suivant la réception du questionnaire, avant que l'on ait jeté à la poubelle le courtier non décaché dans le cadre des travaux normaux d'entretien ménager.

Cependant, les données de ces recherches sont fort peu éclairantes sur l'importance relative des préavis et des appels comme mesures d'incitation.

En général, les résultats des travaux de recherche antérieurs portent à croire que l'insertion d'une enveloppe-réponse affranchie (par comparaison avec une enveloppe-réponse d'affaires) améliore le taux de réponse (Scott 1961; Kanuk et Berenson 1975; Duncan 1979; Harvey 1987; Fox et coll. 1988). Rappelons cependant l'exception digne de mention que constitue une analyse de régression qui porte sur des études de Heberlein et Baumgartner, et qui a montré que l'insertion d'enveloppes-réponse affranchies n'exerçait pas d'effet significatif (1979). Une revue de la documentation effectuée par Armstrong et Luske signale 20 études dans lesquelles des solutions de remplacement aux enveloppes-réponse d'affaires (1987) ont été mises à l'essai. Pour chacune des comparaisons, le taux de réponse absolu suscité par la solution de remplacement était beaucoup plus élevé que celui découlant de la technique habituelle dans 15 cas sur 20, l'augmentation moyenne étant de 9,2 points. On a signalé six études comparant l'utilisation d'enveloppes affranchies à la machine et d'enveloppes affranchies avec un timbre. En moyenne, le taux de réponse découlant de ces dernières était supérieur de 3,4 points à celui des enveloppes affranchies à la machine. Enfin, quatre études reposant sur une pléiade de mesures d'incitation ont montré que les enveloppes affranchies engendraient un taux de réponse supérieur de 2 à 4 points à celui obtenu avec des enveloppes-réponse d'affaires (Dillman 1978).

Les trois mesures d'incitation qui sont testées dans notre enquête font partie des huit principales techniques unifor-mément reconnues dans les rapports de recherche comme des déterminants de l'amélioration du taux de réponse postale. Parmi les autres facteurs, citons les stimulants financiers, les services postaux spéciaux, l'identité de l'organisme enquêteur, la personnalisation de la corres-pondance et l'intérêt du répondant (ou l'importance des questions) (Dillman 1991).

On a jugé que deux de ces huit facteurs, les stimulants financiers et les services postaux spéciaux (p. ex., courtier certifié ou courtier prioritaire avec délai de deux jours), n'étaient pas pratiques dans le cadre d'un recensement touchant plus de 100 millions de ménages. On a estimé qu'un troisième facteur, soit le parrainage du U.S. Bureau of the Census, était souhaitable sur le plan de la stimulation du taux de réponse. Un quatrième facteur, soit l'intérêt du répondant ou l'importance des questions, ne se prêtait pas à une intervention, les questions de l'enquête étant définies à l'avance dans des lois fédérales. Un cinquième facteur, la personnalisation de la correspondance, avait une valeur limitée du fait que les formulaires de recensement s'adressent strictement aux ménages et non à des individus. En examinant les effets individuels et combinés sur les réponses, des préavis, des rappels et des enveloppes-réponse affranchies, nous espérons pouvoir déterminer si l'une ou plusieurs de ces composantes pouvaient se substituer à une autre composante, ce qui nous aurait permis d'améliorer le taux de réponse en réduisant le coût.

Incidence des lettres de préavis, enveloppes-réponse affranchies et cartes de rappel sur les taux de réponse par la poste lors du recensement

DON A. DILLMAN, JON R. CLARK et MICHAEL D. SINCLAIR¹

RÉSUMÉ

Lors d'un recensement national pilote effectué en 1992, la séquence d'envoi d'une lettre de préavis, d'un formulaire de recensement, d'une carte de rappel et d'un questionnaire de remplacement a engendré un taux de réponse global de 63,4%. Ce taux de réponse était considérablement plus élevé que le taux de 49,2% obtenu lors du National Content Test Censur de 1986, dans le cadre duquel on avait également utilisé un questionnaire de remplacement. L'écart a semblé principalement attribuable à la séquence "préavis - formulaire de recensement - rappel" des envois, mais on n'a pas su dans quelle mesure chacun des principaux effets et leurs interactions avaient influé sur les résultats globaux. Le présent article rend compte des résultats du test de mise en oeuvre du Recensement de 1992, qui vérifiait l'influence individuelle et combinée, sur le taux de réponse, de la lettre de préavis, de l'enveloppe-réponse affranchie ainsi que de la carte de rappel. Il s'agissait d'un échantillon national de ménages ($n = 50,000$) interrogés à l'automne 1992. On a utilisé un plan factoriel pour vérifier les huit combinaisons possibles des principaux effets, ainsi que leurs interactions. On a également eu recours à la régression logistique et à des comparaisons multiples pour analyser les résultats du test.

MOTS CLÉS: Enquête par la poste; taux de réponse; comparaisons multiples; régression logistique.

1. INTRODUCTION

Une diminution de 10 points du taux de réponse postale, lequel a glissé de 75% à 65% pour le recensement décennal américain de 1990, a conduit à des recherches visant à améliorer le taux de réponse dans ce genre d'enquête. Chaque point gagné peut faire économiser environ 16 millions de dollars en coûts de dénombrement à domicile (Miskura 1992). Une démarche antérieure nous a appris que des questions simples et quelque peu moins nombreuses que celles qui figuraient dans le questionnaire abrégé de 1990 s'étaient traduites par une amélioration de 8 points du taux de réponse postale (Dillman, Clark et Sinclair 1993). En effet, un formulaire de recensement expérimental comportant ces caractéristiques a été retourné par 71,4% des ménages, alors que 63,4% seulement des ménages d'un groupe de contrôle ayant reçu le formulaire abrégé du Recensement de 1990 l'avaient retourné. Les taux de réponse obtenus avec ces deux formulaires étaient toutefois beaucoup plus élevés que l'avaient été ceux enregistrés à l'occasion de tests similaires effectués des années où il n'y avait pas de recensement. Par exemple, dans le National Content Test de 1986, qui reposait sur un questionnaire équivalant à la version abrégée du formulaire de recensement de 1990, un taux de réponse de 49,2% a été obtenu. On a posé l'hypothèse qu'une partie du taux de réponse élevé qu'on avait constaté dans cette récente expérience était attribuable à une stratégie privilégiant des contacts multiples: lettre de préavis, carte de rappel et questionnaire de remplacement.

Maintes études ont confirmé que le principal déterminant du taux de réponse global aux enquêtes postales est le nombre de contacts (p. ex., Scott 1961; Heberlein et Baumgartner 1978). Les préavis et les rappels se sont tous deux révélés efficaces pour susciter une réponse (p. ex., Kanuk et Berenson 1975; Linsky 1975; Fox et coll. 1988).

1.1 Travaux de recherche antérieurs

Notre article a pour but de rendre compte des résultats du test de mise en oeuvre de 1992, qui vise à déterminer l'influence relative et combinée, sur le taux de réponse, de la lettre de préavis et de la carte de rappel utilisés dans la démarche exposée auparavant (Dillman et coll. 1993). Le test cherchait également à déterminer l'effet produit par l'inclusion d'une enveloppe-réponse affranchie (plutôt qu'une enveloppe-réponse d'affaires) avec le formulaire de recensement envoyé par la poste. Le recensement décennal américain de 1990 nous a obligés à enquêter auprès de plus de 100 millions de ménages. Les seuls coûts d'un tel recensement montrent à quel point il est important de déterminer la mesure dans laquelle ces trois techniques d'incitation à répondre pourraient contribuer à améliorer le taux de réponse des ménages. Si des recherches antérieures laissent deviner l'importance de chacune de ces techniques à cet égard, on dispose de peu de renseignements sur leurs éventuelles interactions. L'étude nous permet de déterminer dans quelle mesure ces trois techniques interagissent ou se complètent lorsqu'elles sont utilisées ensemble.

¹ Don A. Dillman, Washington State University, Pullman, WA, U.S.A.; Jon R. Clark, U.S. Bureau of the Census, Washington DC 20233, U.S.A.; et Michael D. Sinclair, Response Analysis Corp., Princeton, NJ, U.S.A.

ARTHANARI, T.S., et DODGE, Y. (1981). *Mathematical Programming in Statistics*. New York: John Wiley and Sons.

CAUSEY, B.D., COX, L.H., et ERNST, L.R. (1985). Applications of transportation theory to statistical problems. *Journal of the American Statistical Association*, 80, 903-909.

ERNST, L.R. (1986). Maximizing the overlap between surveys when information is incomplete. *European Journal of Operational Research*, 27, 192-200.

ERNST, L.R. (1989). Further Applications of Linear Programming to Sampling Problems. Bureau of the Census, Statistical Research Division, Research Report Series, No. RR-89/05.

ERNST, L.R., et IKEDA, M. (1992a). Modification of the Reduced-Size Transportation Problem for Maximizing Overlap When Primary Sampling Units Are Redefined in the New Design. Bureau of the Census, Statistical Research Division, Technical Note Series, No. TN-91/01.

ERNST, L.R., et IKEDA, M. (1992b). Summary of the Performance of the Maximum Overlap Algorithms for the 1990's Redesign of the Demographic Surveys. Bureau of the Census, Statistical Research Division, Technical Note Series, No. TN-92/01.

ERNST, L.R., et IKEDA, M. (1994). A Reduced-Size Transportation Algorithm for Maximizing the Overlap Between Surveys. Bureau of the Census, Statistical Research Division, Research Report Series, No. RR-93/02.

GLOVER, F., KARNNEY, D., KLINGMAN, D., et NAPIER, A. (1974). A computation study on start procedures, basic change criteria and solution algorithms for transportation problems. *Management Sciences*, 20, 793-813.

KEYFITZ, N. (1951). Sampling with probabilities proportional to size: Adjustment for changes in probabilities. *Journal of the American Statistical Association*, 46, 105-109.

KISH, L., et SCOTT, A. (1971). Retaining units after changing strata and probabilities. *Journal of the American Statistical Association*, 66, 461-470.

PATHAK, P.K., et FAHIMI, M. (1992). Optimal integration of surveys. In *Essays in Honor of D. Basu*. Eds. M. Ghosh, et P.K. Pathak. Hayward, California: Institute of Mathematical Statistics, 208-224.

PERKINS, W.M. (1970). 1970 CPS Redesign: Proposed Method for Deriving Sample PSU Selection Probabilities Within 1970 NSR Stata. Note de service à Joseph Waksberg, Bureau of the Census.

RAJ, D. (1968). *Sampling Theory*. New York: McGraw Hill.

CONCLUSIONS

Pour l'exemple considéré dans les sections 2 et 3, il est possible de procéder à une comparaison valide des différentes méthodes de chevauchement puisque la valeur du chevauchement attendu dans Ernst (1986), soit 1.625, a été facilement calculée à la main. Les valeurs correspondantes du chevauchement pour la méthode à taille réduite et pour la méthode optimale sont 1.725 et 1.735 respectivement.

CPS ou le NCVS puisque le SIPP utilise un plan à deux UPÉ par strate. Vue sous cet angle, la méthode à taille réduite fonctionne mieux que celle proposée par Ernst (1986). Toutefois, comme les stratifications ont été passablement différentes dans ces trois enquêtes, il est permis de douter de la validité d'une telle comparaison.

La méthode de chevauchement à taille réduite présentée dans le présent article répond en pratique à ses deux objectifs principaux. Elle réduit suffisamment la taille du problème de transport, comme le démontrent d'une part la taille du problème de transport dans les formules (3.1) à (3.3), et d'autre part le fait qu'elle a effectivement été utilisée dans le remaniement d'une enquête importante. En outre, cette méthode réalise la réduction de taille tout en donnant un chevauchement presque optimal, à tout le moins dans le cas du SIPP. Elle ne peut être utilisée que lorsque les UPÉ du plan initial sont sélectionnées indépendamment d'une strate à l'autre, mais lorsque cette condition est respectée, nous croyons qu'il s'agit de la meilleure méthode de chevauchement pour les grandes strates.

REMERCIEMENTS

Nous remercions M. Todd Williams pour son aide précieuse à la programmation. Merci également aux critiques et aux rédacteurs pour leurs commentaires constructifs. Les options exprimées dans le présent article sont celles des auteurs et ne sont pas nécessairement partagées par le Bureau of Labor Statistics ni par le Census Bureau.

BIBLIOGRAPHIE

ARAGON, J., et PATHAK, P.K. (1990). An algorithm for optimal integration of two surveys. *Sankhyā: The Indian Journal of Statistics*, 52, 198-203.

Le temps d'exécution de l'algorithme à taille réduite a été relativement court pour la plupart des strates. Nous présentons ci-après les temps machine de la simulation pour la strate finale avec différents nombres d'UPF. Nous avons utilisé un ordinateur Solbourne 5/605. Le nombre médian d'UPF dans une strate, pour le groupe entier de 62 strates, était de 17. La strate la plus grande contenait 68 UPF.

Nous avons également déterminé, pour l'enquête SIPP en conditions réelles, que 41 des 103 strates finales du chevauchement de la méthode modifiée à taille réduite auraient été incompatibles avec la méthode optimale. En effet, selon nos estimations, la taille maximale du problème de transport, en termes de nombre de variables, pour l'exécution du programme, était fixée à 4×10^6 . Le nombre de variables pour la méthode optimale était inférieur à 4×10^6 pour toutes les strates (56) à $n \leq 14$, mais dépassait la limite pour 41 des 47 strates à $n \geq 15$, y compris deux à $n = 15$. La taille maximale du problème de transport avec la méthode optimale pour les 103 strates a été atteinte avec une strate à $n = 46$, pour laquelle il y avait 3.61×10^{12} variables. Par contre, il y avait 1.03×10^6 variables pour la même strate avec la méthode à taille réduite modifiée.

L'efficacité relative du chevauchement de la méthode à taille réduite comparativement à la méthode de Ernst (1986) présente également un intérêt. On pense en général que la méthode à taille réduite devrait produire un chevauchement plus grand dans les cas où les deux méthodes sont utilisables, puisque la méthode à taille réduite tire profit de l'indépendance de strate à strate du plan initial. Toutefois, même si la méthode de Ernst (1986) s'applique aux plans à deux UPF par strate, aucun logiciel n'a jamais été créé au Census Bureau (ni ailleurs à notre connaissance) pour en permettre l'utilisation avec ce type de plan puisqu'il n'y a pas encore eu d'application en conditions réelles pour un tel programme. En conséquence, il n'est pas possible de comparer directement ces deux méthodes avec les mêmes données. On peut toutefois faire une comparaison sommaire à partir des résultats de la méthode de chevauchement à taille réduite pour l'enquête SIPP et des résultats du chevauchement obtenus avec la méthode de Ernst (1986) pour le chevauchement des plans du CPS et du NCVS pour les années 1990 avec leurs contreparties respectives des années 1980. (Les plans des enquêtes CPS et NCVS des années 1980 et 1990 comptent tous une UPF par strate.)

Dans le cas du CPS, la méthode de chevauchement a donné une augmentation moyenne du chevauchement attendu de .26 UPF par strate comparativement au résultat de la sélection indépendante; dans le cas du NCVS, elle a donné une augmentation moyenne du chevauchement attendu de .30 UPF par strate. À titre comparatif, on a obtenu une augmentation de .94 UPF par strate avec la méthode à taille réduite par rapport à la sélection indépendante pour le SIPP. Si les deux méthodes de chevauchement sont également efficaces, on pourrait alors s'attendre à une augmentation du chevauchement par strate à peu près deux fois plus importante pour le SIPP que pour le

avec probabilité des SIPP réalisés dans la région du Midwest au cours des années 1980, en utilisant les données de 1970. Les deux stratification de 1970 ont donné une répartition des UPF sélectionnées avec probabilité en 31 strates, en utilisant différents groupes de variables de stratification. Les stratifications fondées sur les données de 1980 et de 1970 ont été assimilées à des stratifications "initiales" et "finales" aux fins de l'algorithme de chevauchement.

Au cours de l'enquête véritable, tel qu'il est mentionné dans la section 3.1 et comme l'expliquent en détails Ernst et Ikeda (1994), on a utilisé une modification de la méthode à taille réduite pour le chevauchement du plan du SIPP des années 1990 et de celui des années 1980 à cause d'une différence dans la définition des UPF entre les deux décennies. La méthode à taille réduite modifiée a été utilisée pour le chevauchement de 103 strates non auto-représentatives finales (plan des années 1990) du SIPP. Le chevauchement attendu a été calculé pour l'algorithme de chevauchement maximum à taille réduite, pour la sélection indépendante des UPF finales et pour une limite supérieure du chevauchement attendu avec la méthode optimale. On a calculé une limite supérieure au lieu du chevauchement optimal réel parce qu'il est impossible de calculer un chevauchement optimal pour la strate la plus grande. Aux fins de la simulation, la limite supérieure utilisée a été celle mentionnée dans la section 3.3, $\mu_2 + 2\mu_1$, tandis qu'aux fins de l'enquête SIPP, il a fallu utiliser une limite supérieure différente, décrite par Ernst et Ikeda (1994), à cause de la différence des définitions des UPF entre les années 1980 et 1990.

Les résultats des deux stratifications finales de la simulation étaient généralement semblables l'un à l'autre. La combinaison des résultats des deux a donné un chevauchement attendu moyen pour cet ensemble de 62 strates de 1.552, 1.569 et .480 UPF/strate pour la méthode à taille réduite, la limite supérieure du chevauchement optimal et la sélection indépendante respectivement. Pour la mise en oeuvre du SIPP en conditions réelles, les données correspondantes ont été de 1.523, 1.647 et .582 respectivement, alors que les chevauchements attendus correspondant d'UPF pour les 103 strates étaient 156.9, 169.6 et 59.9 respectivement. Ainsi, tant dans les simulations que dans l'enquête SIPP véritable, la méthode à taille réduite a donné des résultats raisonnablement proches de la limite supérieure de la méthode optimale.

Tableau 6
Temps machine pour la méthode à taille réduite

Nombre d'UPF	Temps machine (hh:mm:ss)
18	0:36
37	5:44
49	24:05
68	2:23:43

(1994) examinent les rapports qui existent entre ces paramètres. Nous résumons brièvement ci-après les résultats de leur étude.

Il est admis au départ que $\Omega_I \leq \Omega_r \leq \Omega_O$ pour tout m, m', m'' sont tels que décrit à la section 3.2. En outre, pour le cas qui nous intéresse, $m' = m'' = 2$, les limites inférieures sont établies sur Ω_r , et les limites supérieures sont établies sur Ω_O et $\Omega_I - \Omega_r$.

Par exemple, désignons par μ_2 la probabilité qu'il existe au moins deux éléments dans I , et désignons par μ_1 la probabilité que I soit un ensemble à un seul élément. Soit

$$\lambda = \min\{\pi_i/p_i : i = 1, \dots, n\},$$

$$\min\{\pi_{ij}/p_{ij} : i, j = 1, \dots, n, i \neq j\}, 1\}.$$

Alors $\Omega_O \leq 2\mu_2 + \mu_1$, $\Omega_r \geq \lambda(2\mu_2 + \mu_1/2)$, et $\Omega_O - \Omega_r \leq 2(1 - \lambda)\mu_2 + (1 - \lambda/2)\mu_1$.

Malheureusement, ces limites ne sont pas toujours très strictes. Toutefois, dans certaines circonstances, elles peuvent être utiles. Par exemple, si $\pi_{ij} \geq p_{ij}$ pour tous les i, j et s'il existe une probabilité 1 que I contienne au moins deux éléments, il s'ensuit, compte tenu de ces limites, que $\Omega_r = \Omega_O = 2$.

4. APPLICATION DE LA MÉTHODE À TAILLE RÉDUITE AU SIPP

Nous présentons ci-après les résultats des simulations du chevauchement du SIPP réalisées avant l'enquête, aux fins de la recherche et des essais, ainsi que les résultats du chevauchement du SIPP obtenus en conditions réelles. Pour en savoir plus sur cette question, consulter Ernst et Ikeda (1992b, 1994).

Pour la mise en oeuvre de la méthode de chevauchement à taille réduite, nous avons utilisé un logiciel d'optimisation du flux de coût minimal (minimum cost flow ou MCF) créé par Darwin Kingman et John Mote de la University of Texas, à Austin, qui nous a permis de résoudre le problème de transport requis. Un programme en FORTRAN a été créé pour préparer les données d'entrée et pour traiter les données de sortie du logiciel MCF. Pour tester le logiciel avant la tenue de l'enquête, nous avons utilisé le programme pour effectuer le chevauchement de deux stratifications du SIPP de la région du Midwest fondées sur les données du recensement de 1970 avec la stratification du plan véritable utilisé dans la même région pour le SIPP au cours des années 1980. (À l'époque de la réalisation de ce test, les données du recensement de 1990 n'étaient pas encore disponibles.) Les stratifications de 1970 ont été produites en stratifiant les UPE sélectionnées

liste ordonnée de I . Autrement, les nouvelles probabilités de sélection sont conditionnées sur I lui-même. Ainsi, les nouvelles probabilités de sélection sont conditionnées sur

$f(1), \dots, f(n)$ et des $g_k(1), \dots, g_k(n - k)$ comme dans le cas où $m = 2$. Ensuite, pour chaque $k = 1, \dots, n - 2$, $\ell = 1, \dots, n - k - 1$, un ordonnancement $h_{k\ell}(1), \dots, h_{k\ell}(n - k - \ell)$ de $\{1, \dots, n\} \sim \{f(1), \dots, f(k), g_k(1), \dots, g_k(n - k)\}$ est établi d'une manière semblable à celle utilisée pour $g_k(1), \dots, g_k(n - k)$. Par exemple, en définissant $h_{k\ell}(v)$ pour $v \geq 2$, $p_{f(k),j}^{f(k),j}$ de la définition de $g_k(\ell)$ est remplacé par

$$P(f(k), g_k(\ell), j \in I \text{ et } I \subset (T_k^* \cup g_k(\ell)) \sim$$

$$\{h_{k\ell}(1), \dots, h_{k\ell}(v - 1)\}.$$

Un ordonnancement linéaire des triplets distincts de $\{1, \dots, n\}$ est alors déterminé en représentant chaque triplet par un triplet ordonné unique ayant la forme $(f(k), g_k(\ell), h_{k\ell}(v))$. Un deuxième triplet $(f(k'), g_{k'}(\ell'), h_{k'\ell'}(v'))$ précède le premier si et seulement si $k' < k$, ou $k' = k$ et $\ell' < \ell$, ou $k' = k$ et $\ell' = \ell$ et $v' < v$.

Pour les valeurs de $m \geq 4$, des ordonnancements de m -uplets seraient définis selon une procédure semblable, et les nouvelles probabilités de sélection seraient conditionnées sur $\binom{n}{m} + \binom{n}{m-1} + \dots + \binom{n}{2} + 1$ événements. Pour $m = 1$, les nouvelles probabilités de sélection sont conditionnées sur le premier membre de l'ordonnancement $f(1), \dots, f(n)$ dans I si $I \neq \emptyset$, ou sur \emptyset si $I = \emptyset$.

À noter que si $m > m'$, il est possible qu'au moins une partie des m -uplets ordonnés ne soient pas des sous-ensembles de I , auquel cas ces sous-ensembles seraient exclus de l'ordonnancement et de l'ensemble d'événements sur lesquels les nouvelles probabilités de sélection sont conditionnées. Si aucun m -uplet ne peut être un sous-ensemble de I , alors les nouvelles probabilités de sélection sont conditionnées sur I lui-même.

Il n'est pas nécessaire de limiter les événements initiaux utilisés dans le problème de transport uniquement aux sous-ensembles de I de taille m ou moindre. Par exemple, si $m = 2$ et si $\binom{n}{3} + \binom{n}{2} + n + 1$ est suffisamment petit, on peut alors utiliser une méthode conditionnée sur des sous-ensembles de trois ou moins, ce qui donne un chevauchement attendu généralement plus grand. À l'inverse, si $\binom{n}{n} + \binom{n}{n-1} + \dots + n + 1$ est trop grand, les nouvelles probabilités de sélection peuvent être conditionnées sur des sous-ensembles de I de taille m'' ou moins, où $m'' > m$, en donnant par contre un chevauchement attendu généralement plus petit.

3.3 Rapports entre les chevauchements attendus de la méthode à taille réduite, de la méthode optimale et de la sélection indépendante

Désignons par $\Omega_I, \Omega_r, \Omega_O$ le chevauchement attendu avec la sélection indépendante, la méthode à taille réduite et la méthode optimale respectivement. Ernst et Ikeda

somme penchées. Notons d'abord que $f(1) = 2$ puisque la valeur de π_i/p_i la plus grande s'obtient quand $i = 2$. Trouvons ensuite $g_1(1)$ qui, puisque $f(1) = 2$, est le $f \in \{1, 3\}$ qui donne la valeur maximale de $\pi_{2f}/p_{2f}^{(1)}$. Pour déterminer ce f , posons d'abord $F_\alpha = \{\alpha\}$, $\alpha = 1, 2, 3$, et notons que $T_1^* = \{1, 2, 3\}$. Selon (3.14) avec $\alpha = 2$, $\beta = 1$, il s'ensuit que

$$p_{21}^{(1)} = p_{22}^{(1)}\{1, 2, 3\}p_1^{(1)}\{1, 2, 3\}p_3^{(1)}\{1, 2, 3\} = p_2p_1 \cdot 1 = .45,$$

et on peut de la même façon obtenir que $p_{21}^{(1)} = .525$.

Ainsi, $g_1(1) = 3$, puisque $.5/.525 > .3/.45$. Par conséquent, la première paire de l'ordonnancement est $\{f(1), g_1(1)\} = \{2, 3\}$. Ensuite, $g_1(2) = 1$, puisque 1 est le seul nombre entier qu'il nous reste à utiliser dans l'ordonnancement g_1 , et la seconde paire est de ce fait $\{f(1), g_1(2)\} = \{2, 1\}$. Il n'est pas vraiment nécessaire de déterminer $f(2)$ puisque $\{1, 3\}$ est la seule et donc la dernière paire qui reste, mais à des fins d'illustration, on peut observer que $T_2 = \{1, 3\}$, $p_1^{(2)} = p_1^{(1)}\{1, 3\}p_2^{(1)}\{1, 3\}$ peut observer que $T_2 = \{1, 3\}$, $p_1^{(2)} = p_1^{(1)}\{1, 3\}p_2^{(1)}\{1, 3\}$ même façon $p_3^{(2)} = p_3(1 - p_2) \cdot 1 = .175$. Donc, $f(2) = 3$ puisque $.7/.175 > .5/.15$. En conséquence, $g_2(1) = 1, f(3) = 1$.

3.1.3 Calcul de p_i^* et de c_{ij}^*

Nous expliquons ensuite le calcul des valeurs de p_i^* . Si I_i désigne la paire de nombre entiers $I_i = \{f(k), g_k(\ell)\}$ alors, comme noté antérieurement, $p_i^* = p_{f(k), g_k(\ell)}^{(i)}$. En conséquence, p_i^* peut être calculé à l'aide de la formule (3.14) avec $j = g_k(\ell)$.

Si I_i est un ensemble à un seul élément $\{i\}$ pour un certain $i \in F_\alpha$, alors, comme l'ont établi Ernst et Ikeda (1994),

$$p_i^* = p_{ii}^{(\alpha)}(\{i\}) \prod_{t=1}^n p_t^{(\alpha)}(\emptyset). \quad (3.15)$$

Finalement, si $I_i = \emptyset$, alors

$$p_i^* = \prod_{t=1}^n p_t^{(i)}(\emptyset).$$

Il reste simplement à expliquer le calcul des c_{ij}^* qui, selon (3.5) et (3.6), revient à calculer b_{ii} , $i = 1, \dots, n$.

Pour le calcul de b_{ii} , on observe que

$$b_{ii} = 0 \quad \text{si} \quad I_i = \emptyset,$$

$$= 1 \quad \text{si} \quad I_i = \{v\} \quad \text{et} \quad t = v,$$

$$= 0 \quad \text{si} \quad I_i = \{v\} \quad \text{et} \quad t \neq v,$$

tandis que si $I_i = \{f(k), g_k(\ell)\}$ et $f(k) \in F_\alpha, g_k(\ell) \in F_\beta$, alors $t \in F_\gamma$, alors

$$= 0 \quad \text{si} \quad t \in T_{k\ell} \sim \{g_k(\ell)\}$$

$$= 0 \quad \text{si} \quad t \notin T_{k\ell}^*, \quad (3.17)$$

$$b_{ii} = 1 \quad \text{si} \quad t = f(k) \quad \text{ou} \quad t = g_k(\ell), \quad (3.16)$$

$$\text{et} \quad \gamma = \alpha = \beta, \quad (3.18)$$

$$= \frac{p_{f(k), i}^{f(k), \alpha}(T_{k\ell}^*)}{p_{g_k(\ell), i}^{g_k(\ell), \beta}(T_{k\ell}^*)} \quad \text{si} \quad t \in T_{k\ell} \sim \{g_k(\ell)\} \quad \text{et} \quad \gamma = \alpha \neq \beta, \quad (3.19)$$

$$= \frac{p_{g_k(\ell), i}^{g_k(\ell), \beta}(T_{k\ell}^*)}{p_{f(k), i}^{f(k), \alpha}(T_{k\ell}^*)} \quad \text{si} \quad t \in T_{k\ell} \sim \{g_k(\ell)\} \quad \text{et} \quad \gamma = \beta \neq \alpha, \quad (3.20)$$

$$= \frac{p_{f(k), i}^{f(k), \alpha}(T_{k\ell}^*)}{p_{g_k(\ell), i}^{g_k(\ell), \beta}(T_{k\ell}^*)} \quad \text{si} \quad t \in T_{k\ell} \sim \{g_k(\ell)\} \quad \text{et} \quad \gamma \neq \alpha, \gamma \neq \beta. \quad (3.21)$$

Ernst et Ikeda (1994) démontrent comment les formules (3.16) à (3.21) ont été obtenues.

Lors de la mise en oeuvre du SIPP en conditions réelles, il a fallu modifier la méthode à taille réduite pour réaliser le chevauchement du plan d'échantillonnage du SIPP des années 1990 et de celui des années 1980. Ces modifications se sont avérées nécessaires par suite de changements intervenus dans les définitions des UPB d'une décennie à l'autre. En effet, certaines UPB du plan des années 1990 pouvaient recouper plus d'une UPB du plan des années 1980. Ces modifications sont décrites en détails dans l'article de Ernst et Ikeda (1994).

3.2 Modifications de la méthode à taille réduite pour d'autres plans

En règle générale, on considère n'importe quel plan d'échantillonnage initial sans remise de la forme m' -UPB par strate et n'importe quel plan d'échantillonnage final sans remise de la forme m -UPB par strate, où m', m sont des nombres entiers positifs quelconques. Même si la méthode à taille réduite de la section 3.1 n'a été présentée que pour le cas où $m = m' = 2$, elle peut en fait s'appliquer pour n'importe quelles valeurs de m, m' . Nous présentons brièvement ci-après les modifications nécessaires lorsque $m \neq 2$ ou $m' \neq 2$.

L'utilisation d'une valeur différente de m' ne demande le changement que d'une partie seulement des calculs. Par exemple, si $m = 2$ mais $m' \neq 2$, les calculs de $p_i^{(k)}$, $p_{f(k), i}^{(i)}$ et c_{ij}^* seront différents, mais leurs définitions resteront les mêmes.

Si $m = 3$, peu importe la valeur de m' , on procède à l'ordonnancement de triplets de nombres entiers au lieu de paires dans $\{1, \dots, n\}$. Si I consiste en au moins trois nombres entiers, les nouvelles probabilités de sélection sont conditionnées seulement sur le premier triplet de la

un ordonnancement $f(1), \dots, f(n)$ de $\{1, \dots, n\}$. Ensuite, pour chaque $k = 1, \dots, n - 1$, on établira par récursion un ordonnancement $g_k(1), \dots, g_k(n - k)$ de $\{1, \dots, n\} \sim \{f(1), \dots, f(k)\}$. Un ordonnancement linéaire des paires distinctes de $\{1, \dots, n\}$ pourra alors être déterminé comme suit. Chacune de ces paires peut être représentée uniquement sous la forme d'une paire ordonnée $(f(k), g_k(\ell))$ pour un certain $k \in \{1, \dots, n - 1\}$, $\ell \in \{1, \dots, n - k\}$. Une seconde paire pouvant être représentée sous la forme $(f(k'), g_{k'}(\ell'))$ précède $(f(k), g_k(\ell))$ si et seulement si $k' < k$, ou $k' = k$ et $\ell' < \ell$. À titre d'explication pour l'exemple que nous venons d'examiner, il sera montré plus tard que $f(1) = 2$, $f(2) = 3$, $f(3) = 1$, $g_1(1) = 3$, $g_1(2) = 1$ et $g_2(1) = 1$, et que l'ordonnancement des paires est donc le suivant: $\{2, 3\}$, $\{2, 1\}$, $\{3, 1\}$. L'ordonnancement de f et l'ordonnancement de g_k seront tous les deux effectués pour répondre à l'objectif énoncé au début du présent paragraphe.

Pour obtenir l'ordonnancement $f(1), \dots, f(n)$, on définit de façon récursive $f(k)$, $k = 1, \dots, n$, en choisissant un $f(k) \in T_k$ qui satisfait à

$$\pi_{f(k)}/p_{f(k)}^{(k)} = \max\{\pi_i/p_i^{(k)} : i \in T_k\},$$

où

$$T_1 = \{1, \dots, n\}, \quad T_k = T_{k-1} \sim \{f(k-1)\},$$

$$k = 2, \dots, n, \quad p_i^{(k)} = P(i \in I \text{ et } I \subset T_k),$$

$$k = 1, \dots, n, \quad i \in T_k. \quad (3.7)$$

Puisque $p_i^{(1)} = p_i$, l'ordonnancement que nous venons de définir équivaut à placer d'abord une UPÉ dont la valeur de π_i/p_i^* est la plus élevée. Pour tous les k , $p_{f(k)}^{(k)}$ est la probabilité que $f(k)$ faisait partie de I et que aucun des éléments $k - 1$ précédant $f(k)$ dans l'ordonnancement f ne faisaient partie de I . Ainsi, $p_{f(k)}^{(k)}$ est la probabilité qu'une tentative soit faite de retenir $A_{f(k)}$ dans le nouvel échantillon soit comme premier membre d'une paire ordonnée d'UPÉ de l'échantillon initial, soit comme UPÉ unique de l'échantillon initial dans S . En règle générale, plus $\pi_{f(k)}/p_{f(k)}^{(k)}$ est grand et plus cette tentative risque de porter fruit. Ainsi, la motivation de l'ordonnancement f des UPÉ individuelles est l'analogue de la motivation de l'ordonnancement des paires d'UPÉ dont nous avons déjà parlé.

Il nous reste à expliquer comment calculer $p_i^{(k)}$ pour $k \geq 2$. À cette fin, désignons par r le nombre de strates initiales qui ont des UPÉ en commun avec S , et par F_α , $\alpha = 1, \dots, r$, une partition de $\{1, \dots, n\}$ telle que i et j font partie du même F_α si et seulement si A_i et A_j faisaient partie de la même strate initiale. Soit

$$p_\alpha^i(T) = P(I \cap F_\alpha \subset T), \quad \alpha = 1, \dots, r,$$

$$T \subset \{1, \dots, n\}, \quad (3.8)$$

$$T \subset \{1, \dots, n\}, \quad i \in F_\alpha \cap T. \quad (3.9)$$

$$p_\alpha^i(T) = P(i \in I \text{ et } I \cap F_\alpha \subset T), \quad \alpha = 1, \dots, r,$$

On observe que

$$p_\alpha^i(T) = 1 - \sum_{i \in F_\alpha \sim T} p_i + \sum_{\substack{i, j \in F_\alpha \sim T \\ i < j}} p_{ij}, \quad (3.10)$$

$$p_\alpha^i(T) = p_i - \sum_{j \in F_\alpha \sim T} p_{ij}, \quad (3.11)$$

et finalement, tel que l'ont établi Ernst et Ikeda (1994),

$$p_i^{(k)} = p_\alpha^i(T_k) \prod_{\substack{\ell=1 \\ \ell \neq \alpha}}^r p_\ell^i(T_k), \quad k = 1, \dots, n,$$

$$i \in F_\alpha \cap T_k. \quad (3.12)$$

Ensuite, pour chaque $k = 1, \dots, n - 1$, l'ordonnancement $g_k(\ell)$, $\ell = 1, \dots, n - k$, est défini de façon récursive par le choix de $g_k(\ell) \in T_{k\ell}$ qui satisfait à

$$\pi_{f(k), g_k(\ell)}/p_{f(k), g_k(\ell)}^{(k)} = \max\{\pi_{f(k), j}/p_{f(k), j}^{(k)} : j \in T_{k\ell}\},$$

où

$$T_{k1} = \{1, \dots, n\} \sim \{f(1), \dots, f(k)\},$$

$$T_{k\ell} = T_{k(\ell-1)} \sim \{g_k(\ell-1)\}, \quad \ell = 2, \dots, n - k,$$

$$T_{k\ell}^* = T_{k\ell} \cup \{f(k)\}, \quad \ell = 1, \dots, n - k,$$

$$p_{f(k), j}^{(k)} = P(f(k), j \in I \text{ et } I \subset T_{k\ell}^*),$$

$$\ell = 1, \dots, n - k, \quad j \in T_{k\ell}. \quad (3.13)$$

Noter que $p_{f(k), j}^{(k)}$ désigne ainsi la probabilité conjointe que $f(k)$ soit le premier nombre entier de l'ordonnancement f dans I , qu'aucun des premiers nombres entiers $\ell - 1$ dans l'ordonnancement g_k ne soit dans I et que $j \in I$. En conséquence, $p_{f(k), g_k(\ell)}^{(k)}$ désigne la probabilité que $I^* = \{f(k), g_k(\ell)\}$. Par ailleurs, si $I_\ell = \{f(k), g_k(\ell)\}$, alors $p_i^* = p_{f(k), g_k(\ell)}^{(k)}$, et le choix de $g_k(\ell)$ donne donc la valeur de $\pi_{f(k), g_k(\ell)}/p_i^*$ la plus grande parmi les éléments de $T_{k\ell}$, conformément à l'objectif déjà mentionné de l'ordonnancement des paires d'UPÉ.

Pour calculer $p_{f(k), j}^{(k)}$, Ernst et Ikeda (1994) établissent que si $f(k) \in F_\alpha$, $j \in F_\beta$ alors

$$p_{f(k), j}^{(k)} = p_{f(k), j} \prod_{\substack{\ell=1 \\ \ell \neq \alpha}}^r p_\ell^i(T_{k\ell}^*) \text{ si } \alpha = \beta,$$

$$= p_{f(k), \alpha}^{(k)} (T_{k\ell}^*) p_{j\beta}^{(k)} (T_{k\ell}^*) \prod_{\substack{\ell=1 \\ \ell \neq \alpha, \beta}}^r p_\ell^i(T_{k\ell}^*) \text{ si } \alpha \neq \beta. \quad (3.14)$$

Nous présentons ci-après les calculs utilisés pour obtenir l'ordonnancement de l'exemple sur lequel nous

Ainsi, $c_{11} = b_{11} + b_{12} = 1.6$, et c_{12} , et c_{13} se calculent de la même façon. Pour les six autres rangs du tableau 4, $I_j = I$ et nous savons donc avec certitude quels sont les nombres entiers de I . En conséquence, les c_{ij} de ces six rangs se calculent facilement.

Finalement, nous maximisons le chevauchement attendu (3.1) sous réserve de (3.2) et (3.3), obtenant ainsi les valeurs x_{ij} du tableau 4. Les probabilités conditionnelles $P(N = S_j | I^* = I_j)$ du tableau 5 sont ensuite calculées en divisant chacune des entrées x_{ij} dans le i -ième rang du tableau 4 par p_i^* .

Tableau 5

Probabilités conditionnelles pour la méthode à taille réduite					
i	I_i	1	2	3	
1	{2,3}	0	1/21	20/21	
2	{1,2}	1	0	0	
3	{1,3}	0	1	0	
4	{1}	1	0	0	
5	{2}	1	0	0	
6	{3}	0	1	0	
7	\emptyset	1	0	0	

Le chevauchement attendu pour la méthode à taille réduite est inférieur de .01 à la valeur optimale, c'est-à-dire 1.725 UPÉ. Cet écart par rapport à l'optimalité est dû uniquement au fait que le chevauchement attendu est de 1.6 pour l'événement conjoint $I^* = \{2,3\}$ et $N = \{1,3\}$. Puisque la probabilité de cet événement conjoint est .025 et que la méthode optimale pour cet exemple donne toujours un chevauchement de 2 lorsque l'il existait au moins 2 UPÉ dans l'échantillon initial, l'écart par rapport à l'optimalité est de $.025(2 - 1.6) = .01$.

La méthode à taille réduite ne permet pas de réaliser l'optimalité parce que la paire {2,3} est assortie d'une probabilité de sélection moindre dans le nouvel échantillon que dans l'échantillon initial. Ainsi, tant pour la méthode optimale que pour la méthode à taille réduite, il convient parfois de sélectionner une autre paire (toujours {1,3} pour les deux méthodes de cet exemple) lorsque l'échantillon initial était {2,3}. Les deux méthodes se distinguent en ce que la méthode optimale choisit uniquement {1,3} lorsque $1 \in I$. Avec la méthode à taille réduite, il n'est pas possible d'utiliser l'information concernant $1 \in I$. Ainsi, lorsque $\{2,3\} \subset I$, $1 \in N$ indépendamment de $1 \in I$. C'est là la cause de l'écart par rapport au chevauchement optimal.

3.1.2 Ordonnancement des paires

Nous examinons ci-après comment on parvient généralement à ordonner les paires. Nous désignons à cette fin par $p^{st}, \pi^{st}, s, t = 1, \dots, n, s \neq t$, la probabilité conjointe que $s, t \in I$ et $s, t \in N$, respectivement.

Nous procédons à l'ordonnancement des paires pour les motifs suivants. Si la i -ième paire de l'ordonnancement est $\{s, t\}$ le problème de transport pourra alors conserver cette paire dans le nouvel échantillon lorsque $I^* = I_i$ avec la probabilité conditionnelle $\min\{1, \pi^{st}/p_i^*\}$. (La probabilité de conservation conditionnelle ne peut dépasser ce seuil puisqu'une valeur plus élevée donnerait une probabilité de sélection inconditionnelle supérieure à π^{st} pour la paire du nouveau plan). Ainsi, l'objectif général de l'ordonnancement est de rendre ces probabilités conditionnelles aussi grandes que possible en moyenne pour toutes les paires.

Pour montrer comment l'ordonnancement des paires influence sur le chevauchement attendu, nous examinons l'exemple du tableau 3. Notre méthode d'ordonnancement, comme nous le montrerons plus tard, donne les résultats indiqués et un chevauchement attendu de 1.725 UPÉ. Examinons maintenant une autre solution d'ordonnement pour cet exemple. Soit {1,3} la première paire de l'ordonnancement, {1,2} la seconde et {2,3} la dernière, on obtient $I^* = \{1,3\}$ dans tous les cas où $I = \{1,2,3\}$ ou $I = \{1,3\}$. Ainsi, pour cet ordonnancement, p_i^* est la probabilité que $I^* = \{1,3\}$, et sa valeur est .42. En outre, dans ce cas, $p_j^* = P(I^* = \{2,3\}) = P(I = \{2,3\}) = .21$, alors que les 5 autres colonnes du tableau 3 restent inchangées. L'ordonnancement de rechange donne un tableau de probabilités conditionnelles semblable au tableau 5 sauf dans le rang 1 où les colonnes I_j , $j = 2$ et $j = 3$ deviennent maintenant {1,3}, 10/21 et 11/21 respectivement, et dans le rang 3 où les colonnes correspondantes deviennent {2,3}, 0 et 1 respectivement.

En utilisant la méthode de l'ordonnancement original, il est possible de calculer que le chevauchement attendu pour l'ordonnancement de rechange est inférieur de .055 à la valeur optimale, c'est-à-dire qu'il est de 1.68 UPÉ. La raison pour laquelle cet ordonnancement de rechange donne un chevauchement attendu plus faible est donnée ci-après. En général, l'entrée plus tardive d'une paire dans l'ordonnancement entraîne une valeur p_i^* correspondant plus faible et donc une probabilité de rétention conditionnelle plus élevée lorsque $I^* = I_i$. Ainsi, avec {1,3} au premier rang de l'ordonnancement, $\pi_{13}/p_1^* = 10/21$, ce qui correspond à la probabilité conditionnelle de rétention pour cette paire lorsque $I^* = \{1,3\}$. Par contre, lorsque {1,3} est au troisième rang de l'ordonnancement, $\pi_{13}/p_3^* > 1$ et cette paire est retenue avec certitude. Par ailleurs, la probabilité de rétention conditionnelle de la paire {2,3} lorsque $I^* = \{2,3\}$ augmente aussi pour atteindre 1 lorsque {2,3} passe du premier au troisième rang de l'ordonnancement, mais l'augmentation n'est que de 20/21 et l'ordonnancement original du tableau 3 produit donc une augmentation plus grande que le chevauchement attendu que l'ordonnancement de rechange.

Ainsi, comme le montre cet exemple, l'objectif de l'ordonnancement consiste à placer plus tôt dans l'ordonnancement les paires assorties d'une probabilité de rétention conditionnelle relativement élevée même avec un placement précoce. Pour obtenir l'ordonnancement voulu des paires de nombres entiers, on obtiendra d'abord par récursion

Le problème de transport à taille réduite cherche à conserver les UPB correspondant aux éléments de l'ensemble I^* dans le nouvel échantillon, mais n'utilise pas l'information sur les éléments dans $I \sim I^*$. Ce problème de transport à taille réduite fondé sur l'ensemble de I_i prend la forme décrite ci-après. Désignons par p_i^* la probabilité que $I^* = I_i$, $i = 1, \dots, \binom{n}{2} + n + 1$, et abrégeons $\pi_j^* = P(S_j)$, $j = 1, \dots, \binom{n}{2}$. Pour chaque i, j , la variable x_{ij} correspond à la probabilité conjointe que $I^* = I_i$ et que $N = S_j$, alors que c_{ij} désigne le nombre attendu d'éléments dans $I \cap S_j$ étant donné $I^* = I_i$. Le problème consiste à déterminer le $x_{ij} \geq 0$ qui maximise

$$(3.1) \quad \sum_{i=1}^{\binom{n}{2}+n+1} \sum_{j=1}^{\binom{n}{2}} c_{ij} x_{ij},$$

sous réserve que

$$(3.2) \quad \sum_{j=1}^{\binom{n}{2}} x_{ij} = p_i^*, \quad i = 1, \dots, \binom{n}{2} + n + 1,$$

$$(3.3) \quad \sum_{i=1}^{\binom{n}{2}+n+1} x_{ij} = \pi_j^*, \quad j = 1, \dots, \binom{n}{2}.$$

Une fois obtenues les valeurs optimales de x_{ij} , les probabilités conditionnelles de nouvelle sélection pour S_j , $j = 1, \dots, \binom{n}{2}$, étant donné $I^* = I_i$, sont x_{ij}/p_i^* . À noter que le nombre de variables x_{ij} dans les formules (3.1) à (3.3) est $\binom{n}{2} + n + 1 \times \binom{n}{2}$, comparativement à un maximum de $2^n \times \binom{n}{2}$ dans les formules (2.1) à (2.3). Il nous reste à expliquer la méthode générale utilisée pour ordonner les $\binom{n}{2}$ paires ainsi que les méthodes de calcul des valeurs de p_i^* et de c_{ij} . Avant cela, nous procéderons à une démonstration de la méthode à taille réduite avec l'exemple à deux UPB par strate utilisé dans la section 2 afin d'illustrer la formulation du problème de transport pour la méthode optimale.

L'ordonnement des paires de cet exemple, comme il sera démontré plus tard, est $\{2,3\}$, $\{1,2\}$, $\{1,3\}$. En conséquence, les valeurs I_i sont présentées dans le tableau 3. Noter que si $I = \{1,2,3\}$ ou $I = \{2,3\}$, l'ensemble associé devient alors $I_1 = \{2,3\}$. Pour les six autres possibilités pour I l'ensemble associé est I lui-même.

Par conséquent, en utilisant les valeurs du tableau 1, nous obtenons

$$p_1^* = P(I = \{1,2,3\}) + P(I = \{2,3\}) = .525, \quad (3.4)$$

$p_i^* = P(I_i)$, $i = 2,3$, et $p_i^* = P(I_{i+1})$, $i = 4, \dots, 7$, donnant les valeurs présentées au tableau 3. Puisque $\pi_j^* = P(S_j)$, nous obtenons $\pi_1^* = .30$, $\pi_2^* = .20$, $\pi_3^* = .50$.

Tableau 3
Probabilités des ensembles associés: méthode à taille réduite

I	p_i^*						
	1	2	3	4	5	6	7
I_i	$\{2,3\}$	$\{1,2\}$	$\{1,3\}$	$\{1\}$	$\{2\}$	$\{3\}$	\emptyset
	.525	.135	.105	.045	.09	.07	.03

Les valeurs de c_{ij} pour cet exemple sont présentées au tableau 4. Pour les obtenir, nous avons simplifié les calculs en posant que

$$b_{ii} = P(i \in I \mid I^* = I_i),$$
$$i = 1, \dots, \binom{n}{2} + n + 1, \quad i = 1, \dots, n, \quad (3.5)$$

et en notant que si $S_j = \{s,i\}$, alors

$$c_{ij} = b_{is} + b_{in}. \quad (3.6)$$

Ainsi, le nombre attendu d'éléments dans $I \cap S_j$, sous réserve que $I^* = I_i$, est simplement la somme des probabilités que chacun des deux éléments de S_j soit dans I_i , étant donné $I^* = I_i$. Observer en outre que si le problème de transport pour la méthode optimale connaît la valeur exacte de I et connaît donc avec certitude si chaque élément de S_j est également dans I , tel n'est pas le cas en ce qui concerne la méthode à taille réduite puisque dans ce cas, seul l'ensemble associé I_i est connu. À titre d'illustration, examinons le premier rang du tableau 4. Puisque $I_1 = \{2,3\}$, nous savons que $2 \in I$ et $3 \in I$, et donc que $b_{12} = b_{13} = 1$. Toutefois, nous ne savons pas avec certitude si $1 \in I$ puisque I_1 est l'ensemble associé à la fois pour $I = \{1,2,3\}$ et $I = \{2,3\}$. En fait, en se reportant au tableau 1,

$$b_{11} = \frac{P(I = \{1,2,3\}) + P(I = \{2,3\})}{P(I = \{1,2,3\})} = .6.$$

Tableau 4

Valeurs de c_{ij} et de x_{ij} qui maximisent le chevauchement pour la méthode à taille réduite

i	I_i	c_{ij}			x_{ij}		
		1	2	3	1	2	3
1	$\{2,3\}$	1.6	1.6	2.0	0.000	0.025	0.500
2	$\{1,2\}$	2.0	1.0	1.0	0.000	0.135	0.000
3	$\{1,3\}$	1.0	2.0	1.0	0.000	0.105	0.000
4	$\{1\}$	1.0	1.0	0.0	0.045	0.000	0.000
5	$\{2\}$	1.0	0.0	1.0	0.090	0.000	0.000
6	$\{3\}$	0.0	1.0	1.0	0.000	0.070	0.000
7	\emptyset	0.0	0.0	0.0	0.030	0.000	0.000

taille 2, c'est-à-dire $n^* = \binom{2}{n}$. Toutefois, m^* peut varier largement, $m^* = \binom{2}{n}$ lorsque les UPB de S comprennent une seule strate initiale. La limite supérieure de 2^n sur m^*

est atteinte lorsque toutes les UPB de S étaient dans des strates initiales différentes, comme le montre l'exemple précédent, et comme dans certaines autres situations. Ernst et Ikeda (1994) présentent une expression générale exacte de m^* .

En ce qui concerne le chevauchement des plans d'échantillonnage sans remise à deux UPB par strate, le nombre de variables du problème de transport pour la méthode optimale est m^*n^* , et peut atteindre jusqu'à $2^n\binom{2}{n}$. Si $n = 15$, $2^n\binom{2}{n} = 3,440,640$, ce qui correspond à peu près à la capacité de résolution des ordinateurs que nous avons à notre disposition. Toutefois, $n > 15$ pour presque maximale.

La méthode que nous décrivons ci-après présente les aspects clés suivants: l'ordonnancement particulier des paires d'UPB; la refonte du problème de transport (2.1) à (2.3) aux fins de la méthode à taille réduite; le calcul des probabilités correspondant aux résultats initiaux de la nouvelle formulation; le calcul des coefficients de coût (c_{ij}) de la fonction objective. Dans la section 3.1.1, nous présentons une description détaillée de la méthode à taille réduite, et une nouvelle définition du problème de transport. L'ordonnancement des paires est décrit dans la section 3.1.2. Finalement, nous présentons dans la section 3.1.3 le calcul des probabilités des résultats initiaux et des coefficients de coût.

3. L'ALGORITHME DE LA MÉTHODE À TAILLE RÉDUITE

Dans les travaux antérieurs portant sur la réduction de la taille du problème de transport défini par les formules (2.1) à (2.3), on s'est surtout attaché à réaliser la réduction de la taille tout en maintenant l'optimalité. Par exemple, Aragón et Pathak (1990) conservent l'optimalité et réduisent la taille du problème de 75% lorsque $m^* = n^*$. Malheureusement, lorsque m^* est beaucoup plus grand que n^* , c'est-à-dire lorsque la réduction de la taille serait la plus utile, leur méthode produit une réduction de taille négligable en termes relatifs. Pathak et Fahimi (1992) proposent une généralisation de cette méthode, mais rien n'indique qu'elle puisse aboutir à coup sûr à une réduction de taille sensible en termes relatifs.

Dans la présente section, nous abordons le problème de la réduction de la taille sous un angle différent. Nous sacrifions l'optimalité, à tout le moins en théorie, pour obtenir en retour une réduction du problème de transport qui en autorise une résolution pratique. Nous y arrivons, dans les cas où les plans initial et nouveau comportent tous les deux deux UPB par strate par exemple, en ordonnant toutes les paires d'UPB dans une nouvelle strate puis en conditionnant les nouvelles probabilités de sélection pour tout ensemble initial d'UPB de l'échantillon de taille supérieure à 2 sur la première paire d'UPB de l'ensemble initial plutôt que sur la totalité de cet ensemble initial. Autrement dit, chaque ensemble initial possible d'UPB de l'échantillon comportant plus de deux UPB est combiné à un ensemble de taille 2. Comme nous l'expliquons à la section 4, cette méthode peut en pratique donner un chevauchement presque optimal, en particulier avec un ordonnancement approprié des paires d'UPB tel qu'il est décrits à la section 3.1.2.

La méthode de réduction de la taille s'applique dans tous les cas où les UPB des plans initial et nouveau sont tirées sans remise. Toutefois, nous limiterons notre description détaillée, dans la section 3.1, aux cas où le plan initial et le nouveau plan comportent tous les deux deux UPB par strate. Nous expliquerons ensuite brièvement, dans la section 3.2, les changements nécessaires pour appliquer cette méthode à d'autres types de plans initiaux et nouveaux. Finalement, dans la section 3.3, nous présenterons certains résultats analytiques portant sur les rapports entre le chevauchement attendu de la méthode à taille réduite, de la méthode optimale et de la sélection indépendante. Tout au long de la présente section, nous présumons au départ que les UPB de l'échantillon initial ont toutes été sélectionnées indépendamment d'une strate à l'autre.

3.1 Méthode de réduction de la taille pour deux plans à deux UPB par strate

La méthode que nous décrivons ci-après présente les aspects clés suivants: l'ordonnancement particulier des paires d'UPB; la refonte du problème de transport (2.1) à (2.3) aux fins de la méthode à taille réduite; le calcul des probabilités correspondant aux résultats initiaux de la nouvelle formulation; le calcul des coefficients de coût (c_{ij}) de la fonction objective. Dans la section 3.1.1, nous présentons une description détaillée de la méthode à taille réduite, et une nouvelle définition du problème de transport. L'ordonnancement des paires est décrit dans la section 3.1.2. Finalement, nous présentons dans la section 3.1.3 le calcul des probabilités des résultats initiaux et des coefficients de coût.

3.1.1 Description générale de la méthode

Dans une première étape, les sous-ensembles $\binom{2}{n}$ de $\{1, \dots, n\}$ de taille 2 sont ordonnés selon une méthode que nous décrivons ultérieurement. (Qu'il nous suffise pour l'instant de mentionner que n'importe quel ordonnancement peut servir à réduire la taille du problème de transport. La méthode précise que nous retenons permet de parvenir à ce résultat en conservant la plus grande part possible des gains de la méthode optimale en matière de chevauchement.) Désignons par I_i , $i = 1, \dots, \binom{2}{n}$, le i -ième élément de l'ordonnancement et par $I_{i+1}^{(2)}, \dots, I_{i+n}^{(2)+n}$ les n sous-ensembles à un seul élément. Posons finalement que $I_{i+1}^{(2)} = \emptyset$. Ainsi, les I_i constituent chacun des sous-ensembles de $\{1, \dots, n\}$ de deux éléments ou moins. À chaque possibilité de I correspond un ensemble I^* unique parmi ces $\binom{2}{n} + n + 1$ sous-ensembles, et les nouvelles probabilités de sélection sont conditionnées sur ce I^* plutôt que sur I . Ainsi, les nouvelles probabilités de sélection sont conditionnées sur $\binom{2}{n} + n + 1$ événements, ce plutôt que sur un nombre possible de 2^n événements qui explique la réduction de la taille. Le paramètre I^* est le premier I_i pour lequel $I_i \subset I$. C'est-à-dire que si I consiste en au moins deux nombres entiers, I^* correspondra à la première paire de l'ordonnancement contenue dans I , tandis que si I est un ensemble à un seul élément ou un ensemble vide, on obtiendra alors $I^* = I$.

Désignons par ailleurs par $J_i, i = 1, \dots, m^*$ les valeurs possibles de I et par $S_j, j = 1, \dots, n^*$ les valeurs possibles de N . L'objet de toutes les méthodes de chevauchement est de maximiser le nombre prévu d'UPÉ dans $N \cap I$, tout en préservant les valeurs des $P(S_j)$.

Pour illustrer plus avant certaines de ces notions, supposons un exemple où $n = 3$. Alors, $n^* = 3$ si le nouveau plan compte une ou deux UPÉ par strate, les valeurs de N , c'est-à-dire les S_j , étant $\{1\}, \{2\}, \{3\}$ dans le cas à une UPÉ par strate et $\{1, 2\}, \{1, 3\}, \{2, 3\}$ dans le cas à deux UPÉ par strate. Supposons maintenant que les UPÉ A_1 et A_2 étaient dans une strate initiale et que l'UPÉ A_3 était dans une autre, et qu'il y avait trois UPÉ dans chacune de ces strates initiales. Si le plan initial comptait une UPÉ par strate, alors $m^* = 6$, et les valeurs de I , c'est-à-dire les J_i , étaient $\emptyset, \{1\}, \{2\}, \{3\}, \{1, 3\}, \{2, 3\}$. Si le plan initial comptait deux UPÉ par strate, alors $m^* = 6$, et les valeurs des J_i étaient $\{1\}, \{2\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}$.

Examinons maintenant le problème de transport pour la méthode de chevauchement de Causey, Cox et Ernst (1985). Soit $P(J_i)$ la probabilité que $I = J_i$ et $P(S_j)$ la probabilité que $N = S_j$. En outre, désignons par x_{ij} la variable dénotant la probabilité conjointe de ces deux événements, et désignons par c_{ij} le nombre d'éléments dans $J_i \cap S_j$. Les valeurs des $P(J_i)$, des $P(S_j)$ et des c_{ij} sont connues alors que celles des x_{ij} sont des variables dont il s'agit de déterminer les valeurs optimales. Ainsi, le problème de transport à résoudre consiste à déterminer la valeur de $x_{ij} \geq 0$ qui maximise

$$(2.1) \quad \sum_{i=1}^{m^*} \sum_{j=1}^{n^*} c_{ij} x_{ij}$$

sous réserve que

$$(2.2) \quad \sum_{i=1}^{m^*} x_{ij} = P(J_i), \quad i = 1, \dots, m^*,$$

$$(2.3) \quad \sum_{j=1}^{n^*} x_{ij} = P(S_j), \quad j = 1, \dots, n^*.$$

Noter que dans ce problème de transport, la fonction objective (2.1) est le nombre prévu d'UPÉ de S qui sont dans $N \cap I$. Noter également que les contraintes (2.2) et (2.3) sont requises par les définitions des $P(J_i)$, des $P(S_j)$ et des x_{ij} .

Une fois obtenues les valeurs optimales x_{ij} , la probabilité conditionnelle que $N = S_j$ étant donné $I = J_i$ est donnée par $x_{ij}/P(J_i)$ pour tous les i, j .

Nous présentons ci-après un exemple de l'utilisation des formules (2.1) à (2.3) dans un cas où le plan initial et le plan nouveau sont tous les deux à deux UPÉ par strate et sans remise. Dans cet exemple comme partout ailleurs dans le présent article, nous désignerons par p_i, π_i la probabilité prédéterminée que $i \in I$ et $i \in N$, respectivement.

Supposons une strate finale S avec $n = 3$. Toutes les UPÉ étaient dans des strates différentes. Soit $p_1 = .6, p_2 = .75, p_3 = .7, \pi_1 = .5, \pi_2 = .8, \pi_3 = .7$. Puisque les UPÉ étaient toutes dans des strates initiales différentes, il existe 8 possibilités différentes pour I , dont les probabilités sont présentées dans le tableau 1.

Tableau 1

Probabilités des groupes possibles d'UPÉ de l'échantillon initial							
I	1	2	3	4	5	6	7 8
J_i	$\{1, 2, 3\}$	$\{1, 2\}$	$\{1, 3\}$	$\{2, 3\}$	$\{1\}$	$\{2\}$	$\{3\}$
$P(J_i)$.315	.135	.105	.21	.045	.09	.07 .03

Puisque le nouveau plan d'échantillonnage est à deux UPÉ par strate et sans remise, il existe trois possibilités pour N : les paires $S_1 = \{1, 2\}, S_2 = \{1, 3\}, S_3 = \{2, 3\}$. Ainsi, $P(S_1) = .30, P(S_2) = .20$ et $P(S_3) = .50$. Les valeurs c_{ij} sont présentées dans le tableau 2. En maximisant ensuite (2.1) sous réserve de (2.2) et (2.3) avec les valeurs données des $P(J_i), P(S_j)$ et c_{ij} , on obtient le groupe de valeurs optimales de x_{ij} présentées dans le tableau 2. Finalement, en divisant chacune des entrées x_{ij} d'un rang i du tableau 2 par $P(J_i)$, on obtient un groupe optimal de probabilités conditionnelles $P(S_j | J_i)$. Par exemple, puisque $x_{12} = .025$ et $P(J_1) = .315$, il s'ensuit que $P(S_2 | J_1) = 5/63$.

Tableau 2

Valeurs de c_{ij} et de x_{ij} qui maximisent le chevauchement pour la méthode optimale							
c_{ij}				x_{ij}			
i	1	2	3	1	2	3	
j							

Pour cet exemple, comme on peut le calculer à partir de (2.1) et du tableau 2, le chevauchement attendu en vertu de la procédure optimale est de 1.735 UPÉ. À titre comparatif, le chevauchement attendu si les plans initial et final sont sélectionnés indépendamment est de $p_1 \pi_1 + p_2 \pi_2 + p_3 \pi_3 = 1.39$ UPÉ.

Pour les plans d'échantillonnage sans remise à deux UPÉ par strate, les valeurs possibles de N sont toujours données par les sous-ensembles $\binom{N}{2}$ de $\{1, \dots, n\}$ de

n'était pas sélectionné indépendamment d'une strate à l'autre. En particulier, comme l'avait expliqué Ernst (1986), si l'échantillon initial est lui-même sélectionné par chevauchement avec un plan antérieur, l'hypothèse d'indépendance n'est en général pas valide, ce qui expliquait principalement pourquoi elle n'était pas valide dans ces quatre cas particuliers.

Le second problème posé par la méthode optimale est que le problème du transport risque d'être trop grand pour pouvoir être résolu en pratique. Le Bureau of the Census a également utilisé la programmation linéaire pour le chevauchement du plan du SIPP des années 1990 avec celui des années 1980, deux plans à deux UPF par strate. L'échantillon initial du SIPP a été sélectionné indépendamment d'une strate à l'autre. Toutefois, le problème du transport pour la méthode optimale aurait été trop grand pour autoriser une solution pratique pour plusieurs strates. Cette difficulté se pose du fait que pour chaque nouvelle strate constituée de n UPF, le nombre de variables du problème de transport pour la méthode optimale peut atteindre $2^n \times \binom{2}{n}$. La valeur de n la plus grande pour laquelle un problème de transport comportant un tel nombre de variables peut être résolu avec les moyens informatiques que nous avons à notre disposition était approximativement $n = 15$.

Nous présentons ci-après une formule à taille réduite de la méthode de chevauchement, assimilée à un problème de transport, qui réduit le nombre de variables du SIPP à $\binom{2}{n} + n + 1 \times \binom{2}{n}$, une réduction surprenante pour les valeurs de n modérées à grandes. Cette méthode part de l'hypothèse que l'échantillon initial a été sélectionné indépendamment d'une strate à l'autre; elle n'aurait de ce fait pas pu remplacer la méthode de Ernst (1986) pour le chevauchement des plans du CPS et du NCVS. Cette méthode à taille réduite a été utilisée avec succès pour des strates comportant jusqu'à 68 UPF. À titre de comparaison, pour un $n = 68$, le nombre possible de $2^{68} \times \binom{2}{68}$ variables de la formule non réduite dépasse de loin la taille des problèmes qui peuvent être résolus avec les ordinateurs courants. En outre, même si la réduction de la taille se fait au détriment de l'optimalité, il semble qu'elle donne en pratique des résultats passablement proches des valeurs optimales, comme nous le démontrerons ci-après. La méthode à taille réduite est celle que nous avons utilisée pour le chevauchement du SIPP.

L'examen de la méthode de Causey, Cox et Ernst (1985), présentée dans la section 2, nous sert de toile de fond pour la présentation de la méthode à taille réduite. La méthode à taille réduite est présentée à la section 3. Même si elle peut se prêter à une application générale, nous nous contenterons, pour simplifier la présentation, de ne décrire en détails que le cas où le plan initial et le nouveau plan comportent deux UPF par strate et se font sans remise. Nous présentons également, dans la même section, un petit exemple artificiel de la méthode à taille réduite qui servira à illustrer la méthode et à démontrer que l'ordonnancement des paires d'UPF dans la strate d'un nouveau plan, une étape clé de l'algorithme, influe sur le chevauchement attendu. Nous soulignons également dans cette

section certains résultats analytiques de la comparaison de la méthode à taille réduite et de la méthode optimale. Nous précisons les limites supérieures de la perte du chevauchement attendu découlant de l'utilisation de la méthode à taille réduite au lieu de la méthode optimale. Nous expliquons également que dans certaines situations, cette perte peut s'approcher de deux UPF pour les plans à deux UPF par strate – le pire scénario envisageable. On trouvera dans l'article de Ernst et Ikeda (1994) de plus amples détails ainsi que les preuves des résultats présentés dans cette section, en plus de certains résultats d'autres sections du présent article.

Dans la section 4, nous abordons la question du rendement de la méthode à taille réduite, tant pour le chevauchement du SIPP actuel que pour des simulations artificielles du chevauchement des SIPP antérieurs. Le chevauchement attendu avec cette méthode est comparé à celui de la sélection indépendante des UPF du nouvel échantillon et à la limite supérieure du chevauchement optimal attendu. Les résultats montrent que pour cette application, et contrairement à certains des résultats théoriques décrits dans la section 3, le chevauchement attendu avec la méthode à taille réduite est beaucoup plus grand que si on avait recouru à une sélection indépendante pour choisir les UPF du nouvel échantillon. Ce chevauchement est en fait presque aussi grand que le chevauchement optimal prévu. Nous précisons finalement le temps machine nécessaire pour la méthode à taille réduite en fonction de la taille des strates. Finalement, les conclusions de notre étude sont présentées à la section 5.

2. EXAMEN DE LA MÉTHODE DE CHEVAUchement DE CAUSEY, COX ET ERNST (1985)

La méthode de chevauchement de Causey, Cox et Ernst (1985), comme toutes les méthodes de chevauchement, conditionne d'une certaine manière la sélection des UPF des échantillons de chaque nouvelle strate au choix des UPF de la strate qui faisaient partie de l'échantillon initial. Cette méthode de chevauchement particulière réalise une optimalité réelle en faisant plein usage de cette information et en assimilant la procédure à un problème de transport. Nous en présentons ci-après la description. Précisons tout d'abord certaines des notations qui nous serviront tout au long de notre exposé. Désignons par S une strate du nouveau plan d'échantillonnage. Chacune de ces strates correspond à un problème de chevauchement séparé. Désignons par n le nombre d'UPF comprises dans S , et par A_1, \dots, A_n les UPF en question comprises dans S . Désignons par I le sous-ensemble aléatoire de $\{1, \dots, n\}$ tel que $k \in I$ si et seulement si A_k était présent dans l'échantillon initial, et désignons par N l'ensemble correspondant du nouvel échantillon. Par exemple, si A_2 et A_3 étaient les UPF de S qui faisaient partie de l'échantillon initial et A_1 et A_3 les UPF comprises dans le nouvel échantillon, alors, $I = \{2, 3\}$ et $N = \{1, 3\}$. Désignons par m^* le nombre de valeurs possibles de I et N respectivement.

Un algorithme de transport à taille réduite pour maximiser le chevauchement des enquêtes

LAWRENCE R. ERNST et MICHAEL M. IKEDA¹

RÉSUMÉ

Lorsqu'on remanie un échantillon selon un plan stratifié à plusieurs degrés, il est parfois indiqué de maximiser le nombre d'unités primaires d'échantillonnage retenues dans le nouvel échantillon sans modifier les probabilités de sélection inconditionnelle. Il existe à cette fin une solution optimale qui s'appuie sur la théorie du transport pour une classe très générale de plans d'échantillonnage. Toutefois, à la connaissance des auteurs, cette méthode n'a jamais été utilisée pour la refonte d'une enquête. Cela s'explique en partie du fait que même pour une strate de taille modérée, le problème de transport résultant pourrait être trop grand pour se prêter à une solution pratique. Dans le présent article, nous proposons un algorithme de transport modifié à taille réduite permettant de maximiser le chevauchement, ce qui réduit sensiblement les dimensions du problème. Cette méthode a été utilisée lors du remaniement récent de l'enquête sur le revenu et la participation aux programmes (Survey of Income and Program Participation, ou SIPP). Nous décrivons brièvement le rendement de l'algorithme à taille réduite, d'une part pour le chevauchement du SIPP en conditions réelles, et d'autre part pour des simulations artificielles antérieures du chevauchement du SIPP. Même si cette méthode n'est pas optimale et ne risque de produire, en théorie, que des améliorations négligeables du chevauchement attendu comparativement à la sélection indépendante, elle ouvre en pratique la voie à des améliorations importantes du chevauchement par rapport à la sélection indépendante dans le cas du SIPP et produit généralement un chevauchement presque optimal.

MOTS CLÉS: Programmation linéaire; remaniement de l'échantillon; Survey of Income and Program Participation.

1. INTRODUCTION

Le problème de la maximisation du nombre prévu d'unités primaires d'échantillonnage (UPÉ) retenues dans un échantillon lors d'une enquête avec un plan stratifié pour lequel les UPÉ sont choisies avec une probabilité proportionnelle à la taille a été abordé pour la première fois dans la documentation scientifique par Keyfitz (1951). Typiquement, la maximisation du chevauchement des UPÉ est motivée par la réduction des coûts supplémentaires tels que ceux requis pour la formation des intervieweurs aux fins d'une enquête auprès des ménages, lesquels s'additionnent à chaque changement d'UPÉ. Les méthodes de maximisation du chevauchement n'influent pas sur la probabilité inconditionnelle de sélection d'un ensemble d'UPÉ dans une nouvelle strate, mais conditionnent sa probabilité de sélection d'une manière telle que la probabilité de sélection d'une UPÉ dans le nouvel échantillon est en règle générale plus grande que la probabilité inconditionnelle lorsque cette UPÉ se trouvait dans l'échantillon initial, et moins grande autrement.

Les méthodes de chevauchement sont applicables lorsque l'aménagement soit une nouvelle stratification des UPÉ, soit un changement de leurs probabilités de sélection. Keyfitz (1951) a proposé une méthode optimale, mais uniquement pour les plans à une UPÉ par strate dans le cas spécial où la strate initiale et la nouvelle strate sont identiques et où seules les probabilités de sélection changent. Causey, Cox et Ernst (1985) ont obtenu une

solution optimale du problème du chevauchement sous des conditions très générales en assimilant ce problème à un problème de transport: une forme spéciale de problème de programmation linéaire. Cette méthode n'impose aucune restriction aux changements de définitions des strates ni au nombre d'UPÉ par strate. (Un résultat similaire a été obtenu de façon indépendante par Arthanari et Dodge (1981), même si ces derniers n'ont pas abordé la question des changements dans les définitions des strates. Les deux groupes de chercheurs ont obtenus leurs résultats en généralisant le travail de Raj (1968).) Toutefois, il existe au moins deux autres difficultés avec la méthode de Causey, Cox et Ernst qui peuvent la rendre inutilisable en pratique. La première est examinée par Ernst (1986), et la seconde fait l'objet du présent article.

La première difficulté découle du fait que si l'échantillon initial d'UPÉ n'a pas été sélectionné indépendamment d'une strate à l'autre, l'information nécessaire au calcul de l'ensemble des probabilités conjointes requises par cette méthode risque de ne pas être disponible en pratique. Ernst (1986) a mis au point une méthode de programmation linéaire de rechange à utiliser dans de tels cas. Le Bureau of the Census a eu recours à la programmation linéaire pour réaliser le chevauchement de ses enquêtes démographiques à cinq reprises. À quatre de ces occasions (sélection des plans pour les Current Population Surveys (CPS) et pour les National Crime Victimization Surveys (NCVS) des années 1980 et 1990), la méthode établie par Ernst (1986) a été retenue puisque le plan initial

¹ Lawrence R. Ernst, Chief, Research Group, Office of Compensation and Working Conditions, Bureau of Labor Statistics, Washington, DC 20212, U.S.A.; Michael M. Ikeda, Mathematical Statistician, Statistical Research Division, Bureau of the Census, Washington, DC 20233, U.S.A.

Mesure de l'inégalité du revenu et de l'erreur-type

Mesure	Estimation	Méthode des équations de suppression	Méthode du jackknife avec suppression
Erreur-type			
Médiane	31705	303.3	569.8
Gini	0.3482	0.005	0.005
Faible revenu	0.1980	0.00586	0.00613
Ordonnées de la courbe de Lorenz			
L(0.2)	0.0561	0.00137	0.00175
L(0.4)	0.1745	0.00166	0.00194
L(0.6)	0.3522	0.00246	0.00285
L(0.8)	0.5982	0.00317	0.00393
Part du quintile			
Q(0, 0.2)	0.0561	0.00137	0.00167
Q(0.2, 0.4)	0.1186	0.00159	0.00221
Q(0.4, 0.6)	0.1775	0.00157	0.00282
Q(0.6, 0.8)	0.2461	0.00158	0.00337
Q(0.8, 1.0)	0.4017	0.00395	0.00451

8. SOMMAIRE

La solution au problème consistant à estimer la variance de statistiques complexes comme la mesure de l'inégalité du revenu a longtemps échappé aux statisticiens. On propose souvent une méthode de répétition comme celle du jackknife pour procéder à l'estimation. L'avantage de la linéarisation est qu'elle se prête à de très nombreux plans d'échantillonnage sans nécessiter de calculs laborieux comme d'autres méthodes, par exemple, la méthode bootstrap. Grâce aux fonctions d'estimation et à la décomposition décrite en (2.3), on se rend compte que certains problèmes délicats peuvent être résolus plus facilement. Le présent article aborde aussi la question de l'ordre de grandeur du reste de certaines mesures. Enfin, on peut effectuer une démonstration plus rigoureuse de la méthode pour un plan d'échantillonnage complexe en suivant les recommandations de Shao et Rao (1994).

BIBLIOGRAPHIE

BEACH, C.M., et DAVIDSON, R. (1983). Distribution-free statistical inference with Lorenz curves and income shares. *Review of Economic Studies*, 50, 723-735.

BINDER, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *Revue Internationale de Statistique*, 51, 279-292.

BINDER, D.A. (1991). Use of estimating functions for interval estimation from complex surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 34-42.

BINDER, D.A., et PATAK, Z. (1994). Use of estimating functions for interval estimation from complex surveys. *Journal of the American Statistical Association*, 89, 1035-1044.

FRANCISCO, C.A., et FULLER, W.A. (1986). Estimation of the Distribution function with a complex survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 37-45.

FRANCISCO, C.A., et FULLER, W.A. (1991). Quantile estimation with a complex survey design. *Annals of Statistics*, 19, 454-469.

GLASSER, G.J. (1962). Variance formulas for the mean difference and coefficient of concentration. *Journal of the American Statistical Association*, 57, 648-654.

KOVAR, J.G., RAO, J.N.K., et WU, C.F.J. (1988). Bootstrap and other methods to measure errors in survey estimates. *La Revue Canadienne de Statistique*, 16, 25-45.

NYGÅRD, F., et SANDSTRÖM, A. (1981). *Measuring Income Inequality*. Stockholm: Almqvist and Wiksell International.

RANDELES, R.H. (1982). On the Asymptotic Normality of Statistics with Estimated Parameters. *Annals of Statistics*, 10, 462-474.

RAO, J.N.K., et WU, C.F.J. (1987). Methods for Standard Errors and Confidence Intervals from Survey Data: Some Recent Work. *Actes de la 46ème session, Institut Internationale de Statistique*, 3, 5-19.

RAO, J.N.K., WU, C.F.J., et YUE, K. (1992). Quelques travaux récents sur les méthodes de rééchantillonnage applicables aux enquêtes complexes. *Techniques d'enquêtes*, 18, 225-234.

SÄRNDAAL, C.-E., SWENSSON, B., et WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

SENDLER, W. (1979). On statistical inference in concentration measurement. *Metrika*, 26, 109-122.

SHAO, J., et RAO, J.N.K. (1994). Standard Errors for Low Income Proportions Estimated from Stratified Multi-Stage Samples. *Sankhyā, B*, (à paraître).

SHAO, J. et WU, C.W.J. (1989). A general Theory for Jackknife Variance Estimation. *Annals of Statistics*, 17, 1176-1197.

SHAO, J. (1993). Inferences Based on L-statistics in Survey Problems: Lorenz Curve, Gini Family and Poverty Proportion. *In Proceedings of the Workshop on Statistical Issues in Public Policy Analysis*, Carleton University et l'Université d'Ottawa.

SINGH, M.P., DREW, J.D., GAMBINO, J.G., et MAYDA, F. (1990). *Méthodologie de l'enquête sur la population active du Canada*, n° 71-526 au catalogue, Statistique Canada.

WOODRUFF, R.S. (1952). Confidence intervals for medians and other position measures. *Journal of the American Statistical Association*, 47, 635-646.

De même, l'erreur quadratique moyenne de la part du quantile correspond à

$$\bar{Q}(p_1, p_2) = \frac{1}{I} \sum_{h=1}^s w_{hcl} y_{hcl} I\{\xi_{p_1} < y_{hcl} \leq \xi_{p_2}\}$$

et on en obtient la valeur approximative avec (6.1) en appliquant

$$u_{hc}^* = \frac{1}{I} \sum_{h=1}^s w_{hcl} [(y_{hcl} - \xi_{p_2}) I\{y_{hcl} \leq \xi_{p_2}\} - (y_{hcl} - \xi_{p_1}) I\{y_{hcl} \leq \xi_{p_1}\}]$$

La mesure du faible revenu définie par (1.3) est estimée grâce à

$$\hat{\theta} = F(\hat{M}/2) = \sum_{h=1}^s w_{hcl} I\{y_{hcl} \leq M/2\}.$$

L'erreur quadratique moyenne de la mesure du faible revenu peut être estimée approximativement grâce à l'expression (6.1), où (en partant de l'équation (5.1)):

$$u_{hc}^* = -\frac{f(\hat{M}/2)}{2f(\hat{M})} \sum_{h=1}^s w_{hcl} I\{y_{hcl} \leq M\} - 1/2]$$

$$+ \sum_{h=1}^s w_{hcl} I\{y_{hcl} \leq M/2\} - \hat{\theta}].$$

7. ILLUSTRATION

Nous illustrerons la méthode décrite auparavant au moyen des données sur le revenu familial recueillies dans le cadre de l'Enquête sur les finances des consommateurs (EFC) du Canada. Nous nous servirons du fichier sur le revenu disponible des familles économiques pour l'Ontario en 1988. Par "revenu disponible", on entend le revenu total après impôt mentionné lors du sondage. L'EFC a le même cadre que l'Enquête sur la population active, qui repose elle-même sur un échantillonnage stratifié à plusieurs degrés. Pour en apprendre davantage au sujet du plan d'échantillonnage, voir Singh et coll. (1990). Nous avons estimé la médiane M , le coefficient de Gini G , la mesure du faible revenu θ , les ordonnées de la courbe de Lorenz et la part des quantiles $\bar{Q}(0, .2), \bar{Q}(.2, .4), \bar{Q}(.4, .6), \bar{Q}(.6, .8), \bar{Q}(.8, 1.0)$. Les erreurs-types correspondent ont été obtenues avec la méthode proposée et la méthode du jackknife avec suppression d'une grappe.

Voici une brève description de la dernière méthode, citée aux fins d'illustration. Tout d'abord, on suppose que l'estimation du paramètre inconnu θ est établie par l'expression $\hat{\theta} = \mathcal{L}(\hat{F})$, où \hat{F} représente la fonction de distribution estimée. L'estimation de la fonction de distribution $F(h_j)$ obtenue de l'échantillon après suppression

de la j -ième grappe prélevée de la h -ième strate ($j = 1, \dots, n_h, h = 1, \dots, H$) est

$$F_{(gj)}^{(g)}(y) = \sum_{h=1}^s A_{hcl}(g, j) w_{hcl} I\{y_{hcl} \leq y\}$$

$$\text{où } A_{hcl}(g, j) = \begin{cases} 1, & h \neq g, \\ \frac{n_g}{n_g - 1}, & h = g, c \neq j; \\ 0, & h = g, c = j. \end{cases}$$

Par conséquent, $\hat{\theta}_{(gj)} = \mathcal{L}(F_{(gj)}^{(g)})$ et l'estimateur jackknife, après suppression d'une grappe, de la variance $\hat{\theta} = \mathcal{L}(F)$ est

$$\text{var}_J(\hat{\theta}) = \sum_{h=1}^s \frac{n_g}{n_g - 1} \sum_{j=1}^{J_g} (\hat{\theta}_{(gj)} - \hat{\theta})^2.$$

On sait que l'estimateur jackknife de la variance laisse à désirer avec les quantiles en raison de son manque de convergence (Kovar et coll. 1988). Des résultats plus récents (Shao et Wu 1989; Rao, Wu et Yue 1992) suggèrent que dans certaines conditions, la méthode du jackknife avec suppression du paramètre d ou suppression d'une grappe pourrait avoir les propriétés asymptotiques désirées pour l'estimation de la variance des statistiques non lisses comme les quantiles ou la mesure du faible revenu. Par conséquent, pour d'autres statistiques comme le coefficient de Gini, l'estimateur jackknife de la variance asymptotique est convergent (Shao 1993).

À l'inverse de la méthode du jackknife, la méthode des équations d'estimation n'exige pas d'importants calculs. Elle est simple et explicite, et elle intègre le plan d'échantillonnage. Elle engendre des formules faciles à programmer pour la variance asymptotique, en dépit de leur apparence complexe.

Puisque l'utilisation d'un seul échantillon restreint la comparaison objective de deux méthodes, notre exemple n'a d'autre but que faire ressortir la variation de l'erreur-type telle qu'elle est obtenue avec des équations d'estimation et une méthode de calcul intensif comme celle du jackknife. Les résultats apparaissent sous forme abrégée dans le tableau qui suit. La tendance suivie par la variation de l'erreur-type estimee confirme le caractère généralement classique de la méthode du jackknife. L'écart est attribuable au biais à la hausse introduit par cette méthode dans le cas de la médiane, bien que la méthode du jackknife avec suppression d'une grappe semble préférable à la même méthode avec suppression de la valeur 1. En ce qui concerne la part des quantiles, l'erreur peut s'expliquer en partie par le fait que la part des quantiles supérieurs ne recoupe pas toutes les unités primaires d'échantillonnage mais fonctionne à la manière de classes distinctes dont l'effet sur la méthode du jackknife pourrait être plus marqué que sur la méthode des équations d'estimation.

Un estimateur convergent de l'erreur quadratique moyenne approximative pour la fonction de distribution estimée en y prendrait la forme (6.1), où $u_{hc}^* = \sum w_{hcl} [I\{y_{hcl} \leq y\} - F(y)]$. L'estimation habituelle du quantile de la population finie donne le quantile de l'échantillon

$$\xi_p = \inf\{y_{hcl} : F(y_{hcl}) \geq p\},$$

c'est-à-dire la solution de l'équation d'estimation

$$\sum w_{hcl} [I\{y_{hcl} \leq \xi_p\} - p] = 0.$$

Si on reprend les résultats de (2.5), l'estimateur de l'erreur quadratique moyenne du p -ième quantile correspond à l'équation (6.1) et

$$u_{hc}^* = \frac{1}{2} \sum w_{hcl} [I\{y_{hcl} \leq \xi_p\} - p].$$

Quand on utilise l'expression (5.2) pour estimer la fonction de densité $f(\xi)$, l'estimation de l'erreur quadratique moyenne du quantile ξ_p devient

$$\text{mse}_\alpha(\xi_p) = \left(\frac{D_\alpha(\xi_p)}{2} \right)^2 \quad (6.2)$$

où $D_\alpha(\xi_p) = (h_1 + h_2)/2 = (\xi_U - \xi_L)/2$ représente la demi-longueur de l'intervalle de confiance à $100(1 - \alpha)\%$ pour ξ_p . Dans un plan d'échantillonnage complexe, h_1 et h_2 correspondent respectivement à la solution de

$$\xi_L = \xi_p - h_1 =$$

$$\inf\{y_{hcl} : F(y_{hcl}) \geq p - z_{1-\alpha/2} \sqrt{\text{mse}[\hat{F}(\xi_p)]}\}$$

$$\xi_U = \xi_p + h_2 =$$

$$\inf\{y_{hcl} : F(y_{hcl}) \geq p + z_{1-\alpha/2} \sqrt{\text{mse}[\hat{F}(\xi_p)]}\}.$$

Francisco et Fuller (1991) ont utilisé eux aussi l'estimateur (6.2). En règle générale, l'intervalle de confiance de Woodruff (1952) pour les quantiles individuels explique pourquoi on recourt à (5.2), et par conséquent à (6.2). Francisco et Fuller (1986), puis Rao et Wu (1987) se sont servis de cet intervalle pour trouver des estimateurs de la variance. Bien que l'estimateur dépende du coefficient de confiance, ces auteurs ont montré qu'il est convergent par rapport à l'asymptote à un seuil de signification α . Rao et Wu (1987) ont examiné l'erreur-type des quantiles pour les échantillons en grappes estimés de cette manière. Les résultats qu'ils ont obtenus avec la méthode de Monte Carlo suggèrent qu'un intervalle de confiance de 95% donne de bons résultats quand vient le moment d'établir l'erreur-type. Binder et Patak (1994) ont obtenu un estimateur de la variance analogue par l'approche des équations d'estimation.

L'estimation du coefficient de Gini habituel correspond à la solution de l'équation suivante

$$\sum w_{hcl} \{ [2F(y_{hcl}) - 1] y_{hcl} - G y_{hcl} \} = 0$$

qui se transforme pour donner

$$G = \frac{2}{\mu} \sum w_{hcl} F(y_{hcl}) y_{hcl} - 1$$

$$\text{où } \mu = \sum w_{hcl} y_{hcl}.$$

On peut estimer l'erreur quadratique moyenne pour le coefficient de Gini grâce à l'expression (6.1) en remplaçant u_{hc}^* , défini en (3.1) par son équivalent pour une enquête complexe. Après manipulation algébrique, on parvient à l'expression suivante:

$$u_{hc}^* = \frac{2}{\mu} \sum w_{hcl} \left[A(y_{hcl}) y_{hcl} + B(y_{hcl}) - \frac{\mu}{2} (G + 1) \right]$$

où

$$A(y) = F(y) - \frac{G + 1}{2}$$

et

$$B(y) = \sum w_{hcl} y_{hcl} I\{y_{hcl} \geq y\}.$$

Pour obtenir les ordonnées de la courbe de Lorenz il suffit de résoudre le système d'équations d'estimation suivant:

$$\sum w_{hcl} [I\{y_{hcl} \leq \xi_p\} y_{hcl} - L(p) y_{hcl}] = 0$$

$$\sum w_{hcl} [I\{y_{hcl} \leq \xi_p\} - p] = 0.$$

L'estimation qui en résulte est donnée par:

$$L(p) = \frac{1}{\mu} \sum w_{hcl} y_{hcl} I\{y_{hcl} \leq \xi_p\}.$$

Pour calculer l'estimateur de l'erreur quadratique moyenne des ordonnées de la courbe de Lorenz, il suffit d'utiliser les valeurs u_{hc}^* définies par (6.3) dans (6.1):

$$u_{hc}^* = \frac{1}{\mu} \sum w_{hcl} [(y_{hcl} - \xi_p) I\{y_{hcl} \leq \xi_p\} + p \xi_p - y_{hcl} L(p)]. \quad (6.3)$$

$$0 = \int \left(I \left\{ y \leq \frac{M}{2} \right\} - \theta \right) dF(y)$$

$$\approx \frac{1}{2} (M - M)f\left(\frac{M}{2}\right) - (\hat{\theta} - \theta)$$

$$+ \int \left(I \left\{ y \leq \frac{M}{2} \right\} - \theta \right) dF(y).$$

Si on remplace $M - M$ par le résultat obtenu en (4.1) et résoud $\hat{\theta} - \theta$, on obtient

$$\hat{\theta} - \theta \approx \int u^*(y) dF(y),$$

où

$$u^* = - \frac{f\left(\frac{M}{2}\right)}{2f(M)} \left(I \{ y \leq M \} - \frac{1}{2} \right)$$

$$+ I \left\{ y \leq \frac{M}{2} \right\} - \theta. \quad (5.1)$$

Pour estimer la variance de la mesure estimative du faible revenu à partir de ce résultat, il faut estimer $f(M)$ et $f(M/2)$. À cette fin, on pourrait se servir de

$$f(\xi) = \frac{h}{F\left(\xi + \frac{h}{2}\right) - F\left(\xi - \frac{h}{2}\right)},$$

pourvu que h soit assez faible. Parallèlement, on pourrait procéder au calcul que voici, comme le proposent Francisco et Fuller (1991) pour un problème analogue. Pour une valeur quelconque de ξ , on estime le percentile correspondant $100p$, puis on établit l'intervalle de Woodruff pour le même percentile en résolvant d'abord h_1 et h_2 dans

$$\inf_{h_1} \left[\frac{\int \left[I \{ y \leq \xi - h_1 \} - p \right] dF(y)}{\text{mse} \left\{ \int \left[I \{ y \leq \xi \} - p \right] dF(y) \right\}^{1/2}} \right] \leq -z_{1-\alpha/2},$$

$$\inf_{h_2} \left[\frac{\int \left[I \{ y \leq \xi + h_2 \} - p \right] dF(y)}{\text{mse} \left\{ \int \left[I \{ y \leq \xi \} - p \right] dF(y) \right\}^{1/2}} \right] \geq z_{1-\alpha/2}.$$

6. ESTIMATION DANS LE CADRE D'UNE

ENQUÊTE COMPLEXE

Ce calcul utilise l'équivalent asymptotique de $\hat{\xi} - \xi$ et la somme estimée de $u^*(y)$ donnée par (2.5). Comme on peut le constater, estimer la variance de la mesure du faible revenu s'avère relativement compliqué. La méthode des fonctions d'estimation nous procure toutefois les formules appropriées.

La discussion relative au reste R de la décomposition (2.3) pour la mesure du faible revenu est analogue à celle de l'estimation du quantile (2.5).

$$f(\xi) = \frac{h_1 + h_2}{2z_{1-\alpha/2} \left[\text{mse} \left\{ \int \left[I \{ y \leq \xi \} - p \right] dF(y) \right\}^{1/2} \right]}. \quad (5.2)$$

où $z_{1-\alpha/2}$ est le 100(1 - $\alpha/2$)-ième percentile de la distribution normale type. Il suffit ensuite de calculer

$$\hat{\mu} = \sum_{h_{ci}} w_{h_{ci}} y_{h_{ci}}$$

$$\text{mse}(\hat{\mu}) = \sum_{h_h} \frac{n_h}{n_h} \frac{1}{n_h} \sum_{h_{hc}} (u_{h_{hc}}^* - u_h^*)^2 \quad (6.1)$$

Supposons un échantillonnage stratifié à plusieurs degrés comprenant un grand nombre de strates H et quelques unités d'échantillonnage primaires (grappes), $n_h (\geq 2)$, venant de chaque strate. Dans l'Enquête sur les finances des consommateurs (EFC) du Canada, qui utilise elle-même l'Enquête sur la population active (EPA), par exemple, on compte plusieurs centaines de strates et en moyenne moins de six grappes par strate. Supposons que $w_{h_{ci}}$ est le poids normalisé de la i -ième unité ultime de la h -ième grappe de la h -ième strate. L'estimateur approprié de l'estimateur convergent de l'erreur quadratique moyenne seront

$$\hat{\mu} = \sum_{h_{ci}} w_{h_{ci}} y_{h_{ci}}$$

$$\text{mse}(\hat{\mu}) = \sum_{h_h} \frac{n_h}{n_h} \frac{1}{n_h} \sum_{h_{hc}} (u_{h_{hc}}^* - u_h^*)^2 \quad (6.1)$$

où $u_{h_{hc}}^* = \sum_i w_{h_{ci}}(y_{h_{ci}} - \hat{\mu})$ et $u_h^* = 1/n_h \sum_c u_{h_{hc}}^*$. On se sert de $\sum_s = \sum_h \sum_c \sum_i$ pour désigner la somme sur toutes les unités ultimes de l'échantillon intégrant tous les degrés d'échantillonnage. On suppose que les unités primaires d'échantillonnage sont sélectionnées avec une unité de remplacement.

Ce n'est pas l'efficacité des estimateurs qui nous intéresse mais les propriétés des estimateurs couramment utilisés. Une analyse des estimateurs plus complexes qu'on retrouve dans les ouvrages d'économétrie déborde du cadre de notre travail.

Voici un estimateur de la fonction de distribution de la population finie:

$$F(y) = \sum_s w_{h_{ci}} I \{ y_{h_{ci}} \leq y \}.$$

$$R = \int_2 y D(y) dy - \int (G - G) y dD(y).$$

La première intégrale se ramène à zéro par une intégration par parties, de telle sorte qu'on peut calculer approximativement le reste par l'équation

$$R \approx - (G - G) (\mu_Y - \mu_Y)$$

$$= - (G - G) o_p(n^{-1/2+\delta}), \quad 0 < \delta < 1/2.$$

Par conséquent, on peut dire que $R = o_p(|G - G|)$.

4. ORDONNÉE DE LA COURBE DE LORENZ ET PART DU QUANTILE

L'ordonnée de la courbe de Lorenz a été définie en (1.1). L'estimation exige les deux équations suivantes:

$$n_1(y, L(p)) = I\{y \leq \xi_p\} y - L(p)y, \\ n_2(y) = I\{y \leq \xi_p\} - p.$$

La deuxième équation définit le 100p-ième percentile de la distribution, tandis que la première exprime l'ordonnée de la courbe de Lorenz au point correspondant. Si le reste de (2.3) est négligeable, on obtient l'approximation suivante:

$$0 = \int I\{y \leq \xi_p\} [y dF(y) - L(p) y dF(y)]$$

$$\approx \int_p^{\xi_p} y dF(y) - [L(p) - L(p)] \int_p^{\xi_p} y dF(y)$$

$$+ \int I\{y \leq \xi_p\} - L(p) y dF(y).$$

Le premier terme de l'expression peut faire l'objet d'une approximation plus poussée, comme suit:

$$\int_p^{\xi_p} y dF(y) \approx (\xi_p - \xi_p) \xi_p f(\xi_p).$$

Reprenant (2.5), on constate que

$$\xi_p - \xi_p \approx - \int \frac{1}{f(\xi_p)} [I\{y \leq \xi_p\} - p] dF(y), \quad (4.1)$$

de telle sorte que

Par conséquent, la linéarisation nécessaire pour estimer la variance de l'ordonnée de la courbe de Lorenz s'exprime de la manière suivante:

$$n^*(y) = \frac{1}{\mu_Y} [(y - \xi_p) I\{y \leq \xi_p\} + p \xi_p - y L(p)].$$

Le résultat est identique à celui obtenu par Beach et Davidson (1983) pour la variance et la covariance des ordonnées de la courbe de Lorenz quand les variables aléatoires sont indépendantes mais distribuées de façon identique. Pour estimer la variance, on doit introduire ξ_p et $L(p)$ dans $n^*(y)$.

Trois équations sont nécessaires pour estimer la part du quantile $\tilde{Q}(p_1, p_2)$:

$$n_1(y, \tilde{Q}(p_1, p_2)) = I\{\xi_{p_1} < y \leq \xi_{p_2}\} y - \tilde{Q}(p_1, p_2)y, \\ n_2(y) = I\{y \leq \xi_{p_1}\} - p_1, \\ n_3(y) = I\{y \leq \xi_{p_2}\} - p_2.$$

En reprenant les mêmes arguments qu'auparavant, on parvient à l'équation suivante:

$$\frac{1}{\mu_Y} [(y - \xi_{p_2}) I\{y \leq \xi_{p_2}\} - (y - \xi_{p_1}) I\{y \leq \xi_{p_1}\}]$$

$$+ p_2 \xi_{p_2} - p_1 \xi_{p_1} - y \tilde{Q}(p_1, p_2)].$$

5. MESURE DU FAIBLE REVENU

Cette mesure a été définie en (1.3). Pour procéder à l'estimation, deux équations sont nécessaires:

$$n_1(y, \Theta) = I\left\{y \leq \frac{M}{2}\right\} - \Theta, \\ n_2(y) = I\{y \leq M\} - \frac{1}{2},$$

où M est la médiane de la distribution définie par la deuxième équation, alors que la première définit la mesure du faible revenu par rapport à la médiane. En laissant de côté le reste indiqué en (2.3), on obtient l'approximation suivante:

Dans ce cas, le reste correspond à

$$R = \int [\gamma - \Theta x - (\gamma - \Theta_0 x)] [dF(\gamma) - dF(\gamma)] .$$

Par conséquent,

$$-\frac{\Theta - \Theta_0}{R} = [F(\gamma) - F(\gamma)] x \rightarrow 0,$$

pour toute valeur de γ et valeur finie de x .

Parallèlement, pour les quantiles de population,

$$n = I\{\gamma \leq \Theta_0\} - p,$$

(2.5)

$$n^* = -\frac{1}{I\{\gamma \leq \Theta_0\}} \frac{f(\Theta_0)}{f(\Theta_0)},$$

où $f(\Theta_0)$ correspond à la densité de probabilité au point Θ_0 . Le deuxième élément de (2.5) est le prolongement de la représentation des quantiles de l'échantillon par Bahadur, tel qu'il est décrit par Francisco et Fuller (1991). On se servira des résultats de (2.5) comme ordonnées dans la courbe de Lorenz et comme mesure du faible revenu aux parties 4 et 5.

Après réduction, le reste R devient $R = F(\Theta) - F(\Theta_0) - F(\Theta) + F(\Theta_0)$. Avec un plan d'échantillonnage aléatoire simple, Randles (1982) a montré que $R = o_p(n^{-1/2})$. Lorsque le plan est plus complexe, Shao et Rao (1994) obtiennent un résultat asymptotique analogue, sous réserve de certaines contraintes de régularité: ils établissent d'abord que $\Theta - \Theta_0 = O_p(n^{-1/2})$, puis que $R = o_p(n^{-1/2})$ et enfin que $R = o_p(|\Theta - \Theta_0|)$.

3. COEFFICIENT DE LA FAMILLE DE GINI

Pour le coefficient de la famille de Gini donné en (1.2), on peut se servir de

$$u(\gamma, G_J) = J[F(\gamma)]\gamma - G_J\gamma.$$

L'approche de Binder (1983) n'accepte pas l'estimation de la variance du coefficient de Gini. Au lieu de calculer la variance en scindant le problème en deux – un élément pour l'estimateur du ratio et l'autre pour la variance du numérateur – on peut résoudre celui-ci en une fois au moyen des équations d'estimation.

Laisant de côté le reste de (2.3), on obtient l'approximation suivante:

$$0 = \int \{J[F(\gamma)]\gamma - G_J\gamma\} dF(\gamma)$$

Si on représente la différence $F(\gamma) - F(\gamma)$ par $D(\gamma)$, le reste peut être exprimé comme la somme de deux intégrales

$$R = \int \{2\gamma[F(\gamma) - F(\gamma)] - \gamma(G - G_J)\}$$

$$\times [dF(\gamma) - dF(\gamma)].$$

Examinons le comportement asymptotique du reste R pour le coefficient de Gini habituel G . Le reste est égal à

Lorsque les observations sont indépendantes et sont distribuées de façon identique, la variance correspond à celle décrite par Glasser (1962) et Sandler (1979). Pour estimer la variance, il est nécessaire d'estimer μ_Y , $F(\gamma)$, et G_J dans l'expression n^* .

$$+ J[F(\gamma)]\gamma - G_J\gamma - E\{F(\gamma)J'[F(\gamma)]\gamma\}. \quad (3.1)$$

$$n^* = \frac{1}{I} \left[\int_1^{F(\gamma)} J'(d) F^{-1}(d) dp \right]$$

où

$$G_J - G_J \approx \int n^*(\gamma) dF(\gamma),$$

il s'ensuit que

$$= \int \left[\int_0^\gamma J'[F(\gamma)] \gamma dF(x) \right] dF(\gamma),$$

$$= \int_0^\gamma J'[F(\gamma)] \gamma dF(x) dF(\gamma)$$

$$\int F(\gamma) J'[F(\gamma)] \gamma dF(\gamma)$$

et

$$\approx \int [F(\gamma) - F(\gamma)] J'[F(\gamma)] \gamma dF(\gamma),$$

$$\int \{J[F(\gamma)] - J[F(\gamma)]\} \gamma dF(\gamma)$$

Soient

$$-(G_J - G_J) \int \gamma dF(\gamma) + \int \{J[F(\gamma)]\gamma - G_J\gamma\} dF(\gamma).$$

$$\approx \int \{J[F(\gamma)] - J[F(\gamma)]\} \gamma dF(\gamma)$$

où T_X représente la population totale de X , et où T_X^X et T_X^2 correspondent à l'estimation HT du total des variables X et X^2 , respectivement.

Pareillement, l'équation qui suit permet d'estimer la fonction de distribution

$$NF(y) = \sum_{i \in S} w_i(s) I\{y_i \leq y\},$$

où

$$I\{y_i \leq y\} = \begin{cases} 1 & \text{si } y_i \leq y, \\ 0 & \text{si } y_i > y. \end{cases}$$

Souignons que $F(y)$ converge de façon uniforme et asymptotiquement à $F(y)$, mais n'est pas nécessairement une fonction de distribution au véritable sens du terme, à moins que

$$\sum_{i \in S} w_i(s) = N.$$

En règle générale et sous réserve de certaines conditions de régularité pour les plans d'échantillonnage complexes (Francisco et Fuller 1991),

$$F(y) - F(y) - p \rightarrow 0 \text{ pour toutes les valeurs de } y.$$

En d'autres termes, la fonction de distribution de la population finie $F(y)$ accepte un estimateur convergent $F(y)$. Nous nous servirons plus tard de cette propriété de $F(y)$ pour prouver la convergence des estimateurs de variance linéaires à l'égard de diverses statistiques sur le revenu. Examinons maintenant comment la théorie des équations d'estimation s'applique à l'estimation du paramètre θ_0 d'une population finie, qui est la solution de

$$\int n(y, \theta_0) dF(y) = 0.$$

La valeur de l'équation d'estimation pour θ_0 correspond à la valeur de θ , pour laquelle

$$\int n(y, \theta) dF(y) = 0, \quad (2.2)$$

où $n(y, \theta)$ est une estimation de $n(y, \theta)$.

On peut écrire (2.2) de la façon suivante:

$$0 = \int n(y, \hat{\theta}) dF(y)$$

$$= \int [n(y, \hat{\theta}) - n(y, \theta_0)] dF(y) + \int n(y, \theta_0) dF(y) + R, \quad (2.3)$$

où

$$n^* = \frac{1}{Y} (Y - \theta_0 X),$$

$$n = Y - \theta_0 X,$$

obtient:

Par exemple, quand θ_0 est égal au rapport T_Y/T_X , on

une idée de la variance de $\hat{\theta}$. l'erreur quadratique moyenne à l'estimation pour avoir la variance de $\hat{\theta}$. Puisque $\hat{\theta} - \theta_0$ donne une estimation

Dès qu'on connaît $n^*(y)$, il est assez facile de trouver

$$n^*(y) = - \left[\frac{\partial E\{n(y, \theta)\}}{\partial \theta} \right]_{\theta=\theta_0}^{-1} n(y, \theta_0).$$

où

$$\hat{\theta} - \theta_0 \approx - \left[\frac{\partial E\{n(y, \theta)\}}{\partial \theta} \right]_{\theta=\theta_0}^{-1} \times \int n(y, \theta_0) dF(y) = \int n^*(y) dF(y), \quad (2.4)$$

Grâce aux approximations qui précèdent, on obtient en question, en général.

La plupart des applications n'exigent pas l'estimation de $n(y, \theta)$ au moyen de $\hat{n}(y, \theta)$. Toutefois, dans certains cas (le coefficient de Gini par exemple), on estime la fonction $n(y, \theta)$ pour que la formule (2.2) accepte les cas

asymptotiquement négligeable. avoir pour ordre de grandeur $o_p(|\hat{\theta} - \theta_0|)$ afin d'être

$$= (\hat{\theta} - \theta_0) \left[\frac{\partial E\{n(y, \theta)\}}{\partial \theta} \right]_{\theta=\theta_0}^{-1} + o_p(|\hat{\theta} - \theta_0|).$$

$$\int [n(y, \hat{\theta}) - n(y, \theta_0)] dF(y)$$

où, pour les échantillons importants,

Binder (1983) a étudié le cas où $\hat{n}(y, \theta) = n(y, \theta)$ et

proverons que le reste R est asymptotiquement négligeable pour tous les paramètres examinés.

La décomposition de (2.3) sert de point de départ à tous les calculs de la variance dans le présent article. Nous

$$R = \int [n(y, \hat{\theta}) - n(y, \theta_0)] [dF(y) - dF(y)].$$

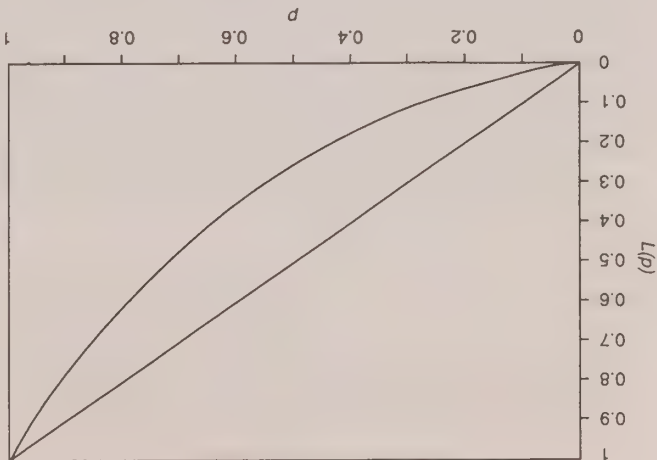


Figure 1. Courbe de Lorenz pour la distribution de Weibull avec $\alpha = 1.6$ comme paramètre de forme.

Le coefficient de Gini détermine le degré d'inégalité de la distribution du revenu. On le définit notamment comme une fonction linéaire de la surface située entre la courbe de Lorenz et l'axe de 45°, normalisée pour rester entre 0 et 1. À la figure 1, le coefficient de Gini est égal à 0.35. La définition formelle du coefficient de Gini (Nygård et Sandström 1981) est la suivante:

$$G = 1 - 2 \int_0^1 L(p) dp = \frac{1}{\mu} \int_0^{\mu} [2F(y) - 1] y dF(y).$$

Nygård et Sandström (1981) donnent le groupe plus général de coefficients de Gini suivant:

$$G_J = \frac{1}{\mu_Y} \int_0^{\mu_Y} J[F(y)] y dF(y), \quad (1.2)$$

où J est une fonction continue et liée. Avec le coefficient de Gini habituel, $J(p) = 2p - 1$. Certains économistes recourent à une autre mesure de l'inégalité du revenu, soit la mesure du faible revenu qui représente la part de la population dont le revenu est inférieur à la moitié du revenu médian de la population. On l'exprime formellement comme suit:

$$\Theta = \int_0^{M/2} dF(y), \quad (1.3a)$$

où M est la médiane définie par

$$\int_0^M dF(y) = \frac{1}{2}. \quad (1.3b)$$

Le paramètre Θ dont nous avons besoin pour ces différentes mesures est la solution à l'équation suivante:

où $u(y, \Theta)$ est le noyau de l'équation d'estimation. Nous reviendrons à la formulation de cette dernière à la partie 2. Les équations d'estimation des mesures qui précèdent et la valeur approximative de leur erreur quadratique moyenne apparaissent aux parties 3, 4 et 5. La partie 6 présente les estimateurs des mesures pour un plan d'échantillonnage complexe. Enfin, la partie 7 illustre la méthode au moyen des données de l'Enquête sur les finances des consommateurs du Canada.

2. APPLICATION DES ÉQUATIONS D'ESTIMATION À UNE POPULATION FINIE

La théorie permettant d'estimer les moyennes et les totaux d'une population finie est très bien développée dans les ouvrages de statistique contemporains. Särndal, Swensson et Wretman (1992) donnent une formule qui englobe la plupart des estimateurs utilisés dans la pratique. Nous passerons ici brièvement en revue cette théorie et montrerons comment il est possible de l'appliquer à des statistiques plus complexes grâce à des équations d'estimation, comme le décrivent Binder (1991) et Binder et Patak (1994). Pour exposer l'idée principale, commençons par examiner l'estimation de la population totale T_Y et la fonction de distribution de la population finie $F(y)$. L'estimation de la population totale se trouve à la base même de l'approche des équations d'estimation de Binder (1991) et de Binder et Patak (1994). Supposons que la population totale de la variable Y soit décrite par

$$T_Y = N \int y dF(y).$$

Souignons que $F(y)$ est une fonction en escalier correspondant à la fonction de distribution de la population finie. Nous utiliserons des estimateurs semblables à

$$\hat{T}_Y = \sum_{i=1}^N w_i(s) y_i = \sum_{i=1}^N w_i(s) X_i, \quad (2.1)$$

où $w_i(s)$ est égal à zéro lorsque la i -ème unité ne fait pas partie de l'échantillon. Par exemple, l'expression (2.1) donne l'estimateur non biaisé d'Horvitz-Thompson (HT) si

$$w_i(s) = \begin{cases} 1/\pi_i, & \text{ies,} \\ 0, & \text{!ies,} \end{cases}$$

ou l'estimateur de régression général si

$$w_i(s) = \begin{cases} [1 + (T_X - F_X) x_i / F_X] / \pi_i, & \text{ies,} \\ 0, & \text{!ies,} \end{cases}$$

Estimation de l'inégalité du revenu d'après les données d'enquête: application de la méthode des équations d'estimation

DAVID A. BINDER et MILORAD S. KOVACEVIC¹

RÉSUMÉ

Les auteurs exposent brièvement quelques-uns des principaux aspects de la théorie des fonctions d'estimation pour les populations finies. Plus précisément, ils abordent la question de l'estimation des moyennes et des totaux et étendent la théorie pertinente aux fonctions d'estimation qu'ils appliquent ensuite au problème de l'estimation de l'inégalité du revenu. Les statistiques qui résultent de l'exercice expriment des fonctions non linéaires des observations, certaines d'entre elles dépendant de l'ordre des observations ou quantiles. Par conséquent, l'erreur quadratique moyenne des estimations ne peut être représentée par une formule simple ni être estimée par les méthodes classiques d'estimation de la variance. Les auteurs montrent que, pour les fonctions d'estimation, il est possible de résoudre ce problème par la méthode de linéarisation de Taylor. Enfin, ils font la démonstration de la méthode proposée en utilisant les données relatives au revenu provenant de l'Enquête sur les finances des consommateurs du Canada et comparent cette méthode à celle du jackknife avec suppression d'une grappe.

MOTS CLÉS: Plan d'enquête complexe; coefficient de Gini; ordonnée de la courbe de Lorenz; mesure du faible revenu, part du quantile.

1. INTRODUCTION

Les ouvrages d'économétrie traitent abondamment des façons de mesurer et d'analyser l'inégalité économique, tant sur le plan théorique que du point de vue des applications, quoiqu'on accorde la préférence aux questions théoriques. On s'est moins intéressé, par contre, à l'estimation de l'inégalité et à l'incidence de la structure des enquêtes par sondage. Les ouvrages d'économétrie pertinents mentionnent rarement l'estimation de la variance, pourtant inévitable dans les inférences statistiques issues de l'estimation de l'inégalité. On contourne habituellement la difficulté en formulant de très fortes hypothèses et en simplifiant à outrance la structure de l'enquête ou les formules servant à obtenir la variance approximative. Nous proposons ici une méthode qui se prête à l'estimation de l'inégalité du revenu et à l'estimation de la variance des statistiques non linéaires qui en dérivent. Cette méthode peut s'appliquer à divers plans d'échantillonnage.

En règle générale, la répartition de la population peut être décrite par sa fonction de distribution cumulative $F(y) = \Pr\{Y \leq y\}$, où Y est la variable aléatoire représentant la sélection d'un sujet au hasard. Nous supposons ici que Y n'a pas de valeur négative. Si X représente le revenu, nous nous intéressons à ses propriétés de distribution, c'est-à-dire à la concentration du revenu, à la part du revenu détenue par certains groupes de la population, à la proportion du faible revenu, etc. La fonction de quantile $\xi(p) = F^{-1}(p) = \inf\{y \mid F(y) \geq p\}$ suscite aussi notre intérêt. La courbe de Lorenz, par exemple, compare le revenu cumulé à la part de la population. La définition officielle de l'ordonnée de la courbe de Lorenz correspond au 100^p-ième percentile de la population est

$$L(p) = \frac{\sum_{Y_i' \in I\{Y_i \leq \xi_p\}} Y_i'}{\sum_{Y_i} Y_i},$$

où I correspond à la population finie et $I\{\cdot\}$ est une fonction indicatrice.

La part du revenu (quantile) est égale au pourcentage du revenu total que partage la population associée au quantile de revenu $[\xi_{p_1}, \xi_{p_2}]$, $p_1 \leq p_2$. Cette valeur est égale à la différence entre deux ordonnées de la courbe de Lorenz

$$\bar{Q}(p_1, p_2) = L(p_2) - L(p_1).$$

La figure 1 illustre graphiquement la courbe de Lorenz pour la distribution de Weibull, avec pour paramètre de forme $\alpha = 1.6$, par rapport à l'axe de 45°. On remarque sur le graphique que la couche défavorisée, qui représente la moitié de la population, reçoit au maximum 25% du revenu total alors que la couche aisée (10% de la population) touche 20% du revenu total disponible.

¹ David A. Binder, Directeur, Division des méthodes d'enquêtes-entreprises, et Milorad S. Kovacevic, méthodologiste principal, Division des méthodes d'enquêtes-ménages, Statistique Canada, immeuble R.H. Coats, Parc Tunney, Ottawa (Ontario), Canada, K1A 0T6.

8. CONCLUSION

Dans un sondage en deux phases, lorsque deux infor-

la population d'une part et l'échantillon issu de la première phase d'autre part, plusieurs stratégies sont possibles lorsqu'on souhaite utiliser l'information auxiliaire pour améliorer l'estimation de totaux.

Deux approches naturelles différentes ont été utilisées pour dériver les estimateurs: une approche assistée par

modèle de régression qui cherche à adapter l'idée de l'estimateur par régression, et une approche calage qui tente d'adapter l'idée du calage. Les estimateurs obtenus par les deux approches peuvent être reliés. On a fait apparaître 3 modélisations sous jacentes alternatives auxquelles peuvent se rattacher l'ensemble des estimateurs obtenus. On obtient ainsi trois classes d'estimateurs. Quelques stratégies de calage auxquelles on pouvait penser ont été éliminées d'emblée comme non pertinentes.

On a montré que les estimateurs d'une même classe, c'est-à-dire rattachés à un même modèle sont asymptoti-

quement équivalents et on donne la forme des variances dérivées dans le cas d'une fonction de calage linéaire, ces résultats restent valables compte tenu des équivalences asymptotiques pour une fonction de calage quelconque.

La forme des variances indiquée, conformément à l'intuition qu'une des classes d'estimateurs (estimateurs 3, 6 et 7) domine l'autre du point de vue de la variance, lorsque l'on se fonde uniquement sur le plan de sondage. Lorsqu'elle s'appuie sur une modélisation de la variable d'intérêt, l'évaluation des stratégies, conduit à privilégier la classe d'estimateurs associée à la modélisation adoptée.

Dans un contexte où l'on souhaite effectuer le redressement d'une enquête, et que l'on souhaite simultanément corriger les biais qu'induit l'utilisation des pondérations brutes et réduire la variance, les conclusions doivent être adaptées. Les modifications introduites dans les pondérations pour corriger les biais sont plus importantes que les corrections pour réduction de variance. Dès lors, on intégrera dans le calage les variables, dès que l'on pense qu'elles interviennent dans la probabilité de sélection et donc participent à la création du biais.

Lorsque les variables auxiliaires sont qualitatives, l'arbitrage entre utilisation a priori et a posteriori de l'information auxiliaire, c'est-à-dire entre, utilisation au niveau de l'échantillonnage ou au niveau du redressement, repose encore sur la distinction entre les deux modèles-

Ces résultats se généralisent sans difficulté au cas d'un sondage à plus de deux phases.

BIBLIOGRAPHIE

- Je remercie chaleureusement Jean-Claude Deville, Louis Meunier et Carl-Erik Särndal pour toutes les suggestions dont ce texte a pu bénéficier.
- CASSEL, C.M., SÄRNDAL, C.-E., et WRETMAN, J.H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite population. *Biometrika*, 63, 615-620.
- DEVILLE, J.-C. (1992). Constrained samples, conditional inference, weighting: three aspects of the utilisation of auxiliary information. *Proceedings of the Workshop on Uses of Auxiliary Information in Surveys*, October 1992, Orebrot.
- DEVILLE, J.-C., et SÄRNDAL, C.-E. (1992). Calibration estimators and generalized raking techniques in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- DEVILLE, J.-C., SÄRNDAL, C.-E., et SAUTORY, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.
- DEVILLE, J.-C., et DUPONT, F. (1993). Calage et redressement de la non-réponse totale. Journées de Méthodologie.
- DUPONT, F. (1994). Redressements alternatifs en présence de plusieurs niveaux d'information auxiliaire. Document de travail de la Direction des Statistiques Démographiques et Sociales, F9409.
- FULLER, W.A. (1975). Regression analysis for sample survey. *Sankhyā C*, 37, 117-132.
- GOURIEROUX, C. (1981). *Théorie des sondages*. Edition Economica Paris.
- ISAKI, C.T., et FULLER, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.
- SÄRNDAL, C.-E. (1980). On π inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika*, 67, 639-650.
- SÄRNDAL, C.-E., et SWENSSON, B. (1987). A general view of estimation for two phases of selection with applications to two-phases sampling and nonresponse. *International Statistical Review*, 55, 279-294.
- SÄRNDAL, C.-E., SWENSSON, B., et WRETMAN, J. (1991). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- WRIGHT, R.L. (1983). Finite population sampling with multivariate auxiliary information. *Journal of the American Statistical Association*, 78, 879-884.

en apportant une réduction de variance (cf. Deville et Dupont 1993). Cependant, asymptotiquement, les corrections de biais à apporter aux poids sont plus importantes que les modifications à apporter pour améliorer les estimateurs. C'est donc le modèle de réponse implicite qui va orienter le choix entre les différents estimateurs :

Le modèle de réponse implicite de la première classe d'estimateurs est $p_i = 1/F(x_{i2}^2 B_2)$

Le modèle de réponse implicite de la deuxième et la troisième classe d'estimateurs est $p_i = 1/F(x_{i2}^2 A_2 + x_{i1}^2 A_1)$.

- Quelque soit le modèle de réponse, l'évaluation des trois classes d'estimateurs sur la base du seul plan de sondage donne encore la préférence à la troisième puisque elle convient pour tous les modèles de réponse.

- Si on évalue les stratégies sur la base d'une modélisation par régression, on n'utilisera la première classe d'estimateurs que si le mécanisme de réponse est bien expliqué par x_2 , c'est-à-dire : $p_i = 1/F(x_{i2}^2 B_2)$. Or, on a vu que la modélisation associée à la première classe d'estimateurs prend son sens lorsque les variables x_1 et x_2 sont très corrélées. Il est donc assez probable dans ce contexte, que la variable x_2 suffise à expliquer le mécanisme de réponse. Dans le cas contraire, il faudrait s'orienter vers la troisième classe d'estimateurs.

La comparaison entre les trois stratégies peut donc être adaptée dans un contexte où on souhaite corriger les biais introduits par les sélection non contrôlées. Les conclusions restent sensiblement les mêmes.

Il est bien sûr possible d'établir selon le même principe, des comparaisons entre stratégies de redressement alternatives dans un contexte de sondages comportant plus de deux phases, et une ou plusieurs sélection non contrôlées.

7. UTILISATION A PRIORI ET A POSTERIORI DE L'INFORMATION AUXILIAIRE

L'estimateur par calage permet d'améliorer a posteriori l'estimation, en réduisant la variance et en corrigeant le biais, comme cela vient d'être mentionné. Toutefois, on peut souhaiter intégrer l'information auxiliaire a priori, au stade du sondage plutôt qu'a posteriori au niveau de l'estimation. On retrouve alors, dans un contexte plus complexe, l'opposition classique entre stratification ou post-stratification, bien connue dans le cas d'un sondage en une phase, lorsque toutes les variables auxiliaires sont qualitatives. Il est possible de transposer les termes de l'arbitrage entre, utilisation de l'information a priori et a posteriori, dans la configuration de sondage et d'information auxiliaire étudiée, lorsque les variables auxiliaires sont qualitatives. En effet, lorsque les variables auxiliaires sont qualitatives, un calage correspond exactement à une poststratification. Nous avons vu précédemment, que pour déterminer la procédure de redressement adéquate, il fallait distinguer deux modélisations de la variable d'intérêt possibles, selon que les informations contenues dans x_1 et x_2 , sont jugées substituables ou complémentaires. Chacune de ces deux

modélisations conduisait alors à une ou des procédures de redressement différentes. De la même façon, ces deux modélisations apparaissent lorsqu'on s'interroge sur la stratégie d'échantillonnage qui permet d'intégrer au mieux l'information auxiliaire :

- Lorsque l'information contenue dans x_1 et celle contenue dans x_2 sont substituables, la modélisation de la variable d'intérêt y est :

$$(1) y_i = x_{i1} b_1 + u_{i1} \text{ et}$$

$$(2) y_i = x_{i2} b_2 + u_{i2} \text{ où le second modèle est de meilleure qualité pour prédire la valeur de } y_i.$$

Nous avons vu que l'utilisation de l'information auxiliaire a posteriori conduit à la stratégie de calage n° 1, soit à la première classe d'estimateurs. Si l'on souhaite tenir compte de l'information auxiliaire au niveau du sondage, il est naturel de proposer, un tirage stratifié en x_1 pour la première phase et un tirage stratifié en x_2 pour la deuxième phase.

Toutefois, le parallèle entre procédure de redressement et procédure d'échantillonnage n'est pas complet : dans un calage, on peut n'utiliser que l'information marginale en x_1 . On réalise alors une poststratification incomplète (Särndal et Deville 1992). En revanche, dans la procédure d'échantillonnage proposée comme alternative a priori, on est amené à utiliser tous les croisements des variables x_1 . L'équivalent a priori d'un calage serait alors un sondage équilibré sur les marges du vecteur de variables x_1 .

- Lorsque l'information contenue dans x_1 et celle contenue dans x_2 sont complémentaires, la modélisation de la variable d'intérêt y est $y_i = x_{i1} b_1 + x_{i2} b_2 + u_i$. Nous avons vu que dans ce cas l'utilisation de l'information auxiliaire a posteriori conduisait aux stratégies de calage 2, 3 et 4 aux classes d'estimateurs 2 et 3. Si l'on souhaite tenir compte de l'information auxiliaire au niveau du sondage, il est naturel de proposer, un tirage stratifié en x_1 pour la première phase et un tirage stratifié en x_2 pour la deuxième phase.

De la même façon que précédemment, il n'y a pas un parallèle exact entre les procédures a priori et a posteriori puisque l'utilisation de l'information a priori mobilise tous les croisements entre les variables x_1 et x_2 .

Ainsi, il est possible d'effectuer un arbitrage entre intégrer l'information a priori ou a posteriori, voire d'optimiser le plan de sondage, lorsque les variables auxiliaires sont qualitatives. Les termes de l'arbitrage sont les mêmes que dans un sondage en une phase avec un seul niveau d'information. Il s'y ajoute la multiplicité des strates créées par les croisements de x_1 et x_2 dans le cas où la modélisation utilisée est $y_i = x_{i1} b_1 + x_{i2} b_2 + u_i$ qui renforce les avantages de l'utilisation a posteriori de l'information. Lorsque les variables auxiliaires sont quantitatives, l'arbitrage passe par leur transformation en variables qualitatives, l'extension correcte ne pouvant être réalisée qu'en utilisant le parallèle entre calage et sondage équilibré (cf. Deville 1992).

$$V(Y_4) =$$

$$\left(\sum_{i,j \in s} \sum_{a \in U} \frac{\Delta_1^2}{\Delta_2^2} \frac{\pi_{aj} \pi_{ai}}{\pi_j \pi_i} \right) + \left(\sum_{i,j \in s} \frac{\Delta_1^2}{\Delta_2^2} \frac{\pi_{aj} \pi_{ai}}{\pi_j \pi_i} \frac{\bar{u}_2 \bar{u}_2}{\bar{u}_1 \bar{u}_1} \right),$$

avec:

$$\bar{u}_{1i} = y_i - x'_{1i} b_1,$$

$$\bar{u}_{2i} = y_i - x'_{2i} b_2.$$

5.2 Estimateurs $\hat{Y}_2 = \hat{Y}_5$: modèle

$$y_i = x'_{1i} a_1 + x'_{2i} a_2 + u_i$$

On montre (voir Dupont 1994) facilement que:

$$VA(\hat{Y}_2) = VA(\hat{Y}_5) \equiv$$

$$\left(\sum_{i,j \in U} \Delta_1^2 \frac{v_i v_j}{v_i v_j} \right) + \left(E_s \sum_{i,j \in s_a} \Delta_2^2 \frac{u_i u_j}{u_i u_j} \right),$$

avec:

$$v_i = y_i - x'_{1i} a_1,$$

$$u_i = y_i - x'_{1i} a_1 - x'_{2i} a_2$$

$$a = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} \sum_{i \in U} x_{1i} x'_{1i} & \sum_{i \in U} x_{1i} x'_{1i} x_{2i} x'_{2i} \\ \sum_{i \in U} x_{2i} x'_{2i} & \sum_{i \in U} x_{2i} x'_{2i} \end{pmatrix} \begin{pmatrix} \sum_{i \in U} x_{1i} y_i \\ \sum_{i \in U} x_{2i} y_i \end{pmatrix}.$$

On en déduit que:

$$V(\hat{Y}_2) = V(\hat{Y}_5) =$$

$$\left(\sum_{i,j \in s} \frac{\Delta_1^2}{\Delta_2^2} \frac{\pi_{ij} \pi_{aj}}{\pi_i \pi_j} \right) + \left(\sum_{i,j \in s} \frac{\Delta_2^2}{\Delta_2^2} \frac{\pi_{aj} \pi_{ai}}{\pi_i \pi_j} \frac{\bar{u}_i \bar{u}_j}{\bar{u}_i \bar{u}_j} \right),$$

avec:

$$\bar{v}_i = y_i - x'_{1i} \hat{a}_1,$$

$$\bar{u}_i = y_i - x'_{1i} \hat{a}_1 - x'_{2i} \hat{a}_2.$$

On retrouve dans cette écriture que par construction de $\hat{Y}_2 - \hat{Y}_5$, x_1 réduit la variance apportée par la première phase et x_2 sont utilisés simultanément pour réduire la variance apportée par la deuxième phase.

5.3 Estimateurs \hat{Y}_3 , \hat{Y}_6 et \hat{Y}_7 : modèle

$$y_i = x'_{1i} c_1 + M_{x_1}(x_2)' c_2 + u_i$$

On montre que $VA(\hat{Y}_6) = VA(\hat{Y}_7) = VA(\hat{Y}_3)$. Ainsi:

$$VA(\hat{Y}_3) = VA(\hat{Y}_6) = VA(\hat{Y}_7) \equiv$$

$$\left(\sum_{i,j \in U} \Delta_1^2 \frac{\pi_i \pi_j}{u_i u_j} \right) + \left(E_s \sum_{i,j \in s_a} \Delta_2^2 \frac{\pi_i \pi_{aj} \pi_j \pi_{aj}}{u_i u_j} \right),$$

On en déduit les trois estimateurs de variance qui diffèrent en raison de coefficients estimés différents:

$$V(\hat{Y}_3) =$$

$$\left(\sum_{i,j \in s} \frac{\Delta_1^2}{\Delta_2^2} \frac{\pi_{ij} \pi_{aj}}{\pi_i \pi_j} \right) + \left(\sum_{i,j \in s} \frac{\Delta_2^2}{\Delta_2^2} \frac{\pi_{aj} \pi_{ai}}{\pi_i \pi_j} \frac{\bar{u}_i \bar{u}_j}{\bar{u}_i \bar{u}_j} \right),$$

$$\bar{u}_{1i} = y_i - x'_{1i} \hat{c}_1,$$

$$\bar{u}_i = y_i - x'_{1i} \hat{c}_1 - M_{x_1} x'_{2i} \hat{c}_2,$$

$$V(\hat{Y}_6) =$$

$$\left(\sum_{i,j \in s} \frac{\Delta_1^2}{\Delta_2^2} \frac{\pi_{ij} \pi_{aj}}{\pi_i \pi_j} \right) + \left(\sum_{i,j \in s} \frac{\Delta_2^2}{\Delta_2^2} \frac{\pi_{aj} \pi_{ai}}{\pi_i \pi_j} \frac{\bar{u}_i \bar{u}_j}{\bar{u}_i \bar{u}_j} \right),$$

$$\bar{u}_{1i} = y_i - x'_{1i} \hat{c}_1,$$

$$\bar{u}_i = y_i - x'_{1i} \hat{a}_1 - x'_{2i} \hat{a}_2,$$

$$V(\hat{Y}_7) =$$

$$\left(\sum_{i,j \in s} \frac{\Delta_1^2}{\Delta_2^2} \frac{\pi_{ij} \pi_{aj}}{\pi_i \pi_j} \right) + \left(\sum_{i,j \in s} \frac{\Delta_2^2}{\Delta_2^2} \frac{\pi_{aj} \pi_{ai}}{\pi_i \pi_j} \frac{\bar{u}_i \bar{u}_j}{\bar{u}_i \bar{u}_j} \right),$$

$$\bar{u}_{1i} = y_i - x'_{1i} \hat{c}_1^*,$$

$$\bar{u}_i = y_i - x'_{1i} \hat{a}_1^* - x'_{2i} \hat{a}_2^*,$$

$$\hat{c}_1 = \hat{a}_1^* + \left(\sum_{i,j \in s_a} \frac{\pi_a}{x_1 x_2} \right)^{-1} \left(\sum_{i,j \in s_a} \frac{\pi_a}{x_1 x_2} \right) \hat{a}_2^*.$$

On retrouve que par construction de \hat{Y}_3 , \hat{Y}_6 et \hat{Y}_7 , x_1 est utilisée pour réduire *au maximum* la variance apportée par la première phase et x_2 permet de réduire la variance apportée par la deuxième phase.

6. CHOIX D'ESTIMATEURS EN PRÉSENCE DE BIAIS DE SÉLECTION

En pratique, lorsqu'on effectue le redressement d'une enquête, il est fréquent que l'on souhaite non seulement améliorer l'estimation, mais aussi, et surtout corriger les biais qu'introduisent les sélections non contrôlées d'individus, comme la non réponse.

Étudions le cas d'un sondage en deux phases, dont la deuxième phase correspond à la non réponse totale. Les poids π_i du sondage 2^{ème} phase sont alors inconnus. Le calage de s va permettre d'estimer ces probabilités, tout

En effet, posons $w = y - x_1' \hat{a}_1^* - x_2' \hat{a}_2^*$. Alors $\hat{Y}_2^* = \hat{Y}_6 + [\hat{W}^* - \hat{W}]$. Or, asymptotiquement $[\hat{W}^* - \hat{W}]$ est un infiniment petit négligeable devant \hat{Y}_6 :

$$[\hat{W}^* - \hat{W}] = \left(\sum^s \frac{\pi_a \pi}{w x_1'} \right) \left(\sum^{s_a} \frac{\pi_a}{x_1' x_1'} \right)^{-1} [X_1 - \hat{X}_1],$$

et

$$\left(\sum^s \frac{\pi_a \pi}{w x_1'} \right) \text{ tend vers zéro et } [X_1 - \hat{X}_1] = O\left(\frac{1}{\sqrt{m}}\right).$$

On obtient finalement: $\hat{Y}_7 \equiv \hat{Y}_6$.

En conclusion, lorsque la fonction de calage est exponentielle l'estimateur \hat{Y}_7 coïncide exactement avec le précédent. Lorsque F est linéaire, \hat{Y}_7 est proche du précédent et correspond donc encore à l'approche modèle de régression dans le cas où l'information contenue dans x_1 est jugée complémentaire à l'information contenue dans x_2 pour estimer y et où on utilise la décomposition $y_i = x_{i1}' c_1 + M_{x_1}(x_{i2})' c_2 + u_i$.

Conclusion: les trois classes d'estimateurs

On vient de voir que les quatre stratégies dérivées d'une approche par calage pouvaient être associées à une modalisation par régression. On obtient ainsi trois classes d'estimateur qui sont:

$Y_4 \equiv Y_1$ associées aux modèles

$$(1) \quad y_i = x_{i1}' b_1 + u_i,$$

et

$$(2) \quad y_i = x_{i2}' b_2 + u_i$$

$Y_5 = Y_2$ associé au modèle

$$y_i = x_{i1}' a_1 + x_{i2}' a_2 + u_i$$

$Y_6 \equiv Y_3$ et $Y_7 \equiv Y_3$ associés au modèle

$$y_i = x_{i1}' c_1 + M_{x_1}(x_{i2})' c_2 + u_i.$$

L'approximation \equiv qui indique que les estimateurs sont rattachés au même modèle de régression prend tout son sens lorsqu'on s'intéresse au calcul de variance de ces différents estimateurs. Les estimateurs qui sont rattachés au même modèle de régression ont en effet la même variance asymptotique.

5. ESTIMATION DE VARIANCES

Étudions les variances des différents estimateurs Y_1, \dots, Y_7 définis précédemment. VA désigne la variance asymptotique d'un estimateur qui est obtenue lorsque N, n et m tendent vers l'infini dans un rapport constant.

• Estimateur Y_1

5.1 Estimateur Y_1 et Y_4 : modèle $y_i = x_{i1}' b_1 + u_i$ et (2) $y_i = x_{i2}' b_2 + u_i$.

La variance de cette estimateur et son estimation sont données dans l'ouvrage de Särndal, Swensson et Wretman (1991). La variance se décompose en deux termes qui mesurent la part de variance due respectivement à la première et à la deuxième phase du sondage.

$$VA(Y_1) = \left(\sum_{i,j \in U} \Delta_{ij} \frac{\pi_i \pi_j}{u_{i1} u_{j1}} \right) + \left(E^{s_a} \sum_{i,j \in s_a} \Delta_{ij} \frac{\pi_i \pi_j}{u_{i2} u_{j2}} \right).$$

avec: $\Delta_{ij}^2 = \pi_{ij} - \pi_i \pi_j$,

$$\Delta_{ij}^1 = \pi_{aj} - \pi_{ai} \pi_{aj},$$

$$u_{i1} = y_i - x_{i1}' b_1,$$

$$u_{i2} = y_i - x_{i2}' b_2,$$

$$b_1 = \left(\sum_{i \in U} x_{i1} x_{i1}' \right)^{-1} \left(\sum_{i \in U} x_{i1} y_i \right),$$

$$b_2 = \left(\sum_{i \in U} x_{i2} x_{i2}' \right)^{-1} \left(\sum_{i \in U} x_{i2} y_i \right).$$

Ainsi l'estimateur de variance se décompose aussi en deux termes qui estiment la part de variance relative à chacune des phases de tirage. On retrouve que par construction de Y_1, x_1 permet de réduire la variance apportée par la première phase et x_2 permet de réduire la variance apportée par la deuxième phase

$$V(Y_1) =$$

$$\left(\sum_{i,j \in U} \Delta_{ij}^1 \frac{\pi_{ij} \pi_{aj}}{u_{i1} u_{j1}} \right) + \left(\sum_{i,j \in s_a} \Delta_{ij}^2 \frac{\pi_{ij} \pi_{aj}}{u_{i2} u_{j2}} \right),$$

1^{ère} phase 2^{ème} phase

avec:

$$u_{i1} = y_i - x_{i1}' b_1,$$

$$u_{i2} = y_i - x_{i2}' b_2.$$

Une telle décomposition repose sur l'écriture $V(Y_1) = V(E[Y_1 | s_a]) + E[V(Y_1 | s_a)]$ qui interviendra pour tous les autres estimateurs.

• Estimateur Y_4

Les termes du développement au premier ordre en $1/\sqrt{m}$ de Y_1 et Y_4 coïncident exactement. On peut donc donner un sens plus précis à l'écriture $Y_4 \equiv Y_1$. On en déduit que $VA(Y_1) = VA(Y_4)$. Ainsi:

Or Y_1 se réécrit:

$$Y_1 = [X_1' b_1] + [X_2^* ' b_2 - X_1' b_1] + [\hat{Y} - \hat{X}_2^* ' b_2].$$

On retrouve donc un estimateur semblable à l'estimateur Y_1 issu de l'approche modèle dans le cas où l'information contenue dans x_1 est jugée substituable à l'information contenue dans x_2 pour estimer y et de moins bonne qualité. Les différences entre Y_1 et Y_4 portent sur les points suivants:

1. B_2 est estimé en intégrant les modifications du calage en x_1 contrairement à b_2 .

2. L'estimation $B_1 = \left(\sum_{s_a} \frac{\pi_a}{x_1 x_1'} \right)^{-1} \left(\sum_s \frac{\pi_a \pi}{x_1 y} \right)$ de B_1 , est effectuée en partie sur s_a contrairement à b_1 .

3. Enfin on utilise les poids redressés $F(x_1 \beta_1) / \pi_a \pi$ dans les sommes en x_2 sur s et sur S_a dans Y_4 contrairement à ce que l'on fait pour Y_1 : l'estimation en x_2 est améliorée par la connaissance de x_1 .

La modélisation sous-jacente est donc bien ici: (1) $y_i = x_1' b_1 + u_i$ et (2) $y_i = x_2' b_2 + u_i$ dont on pense que le second est a priori meilleur pour prédire la valeur de y_i .

Stratégie 2

On obtient des poids

$$w_5^i = \frac{F(x_1' \alpha_1 + x_2' \alpha_2)}{\pi_{a1} \pi_i},$$

l'estimateur associé se réécrit:

$$Y_5 = [X_1' a_1] + [X_2' a_2] + [\hat{Y} - \hat{X}_1' a_1 - \hat{X}_2' a_2].$$

On retrouve donc exactement l'estimateur Y_2 proposé dans l'approche modèle de régression dans le cas où l'information contenue dans x_1 est jugée complémentaire à l'information contenue dans x_2 pour estimer y . Le modèle sous-jacent est bien ici $y_i = x_{11} a_1 + x_{12} a_2 + u_i$.

Stratégie 3

On obtient des poids

$$w_6^i = \frac{F(x_1' \gamma_1 + x_2' \gamma_2)}{\pi_{a1} \pi_i},$$

l'estimateur associé se réécrit:

$$Y_6 = \hat{Y} + [X_1 - \hat{X}_1]' a_1 + [X_2^* - \hat{X}_2]' a_2$$

soit:

$$Y_6 = [X_1' a_1] + [X_2^* ' a_2] + [\hat{Y} - \hat{X}_1' a_1 - \hat{X}_2' a_2].$$

Y_7 et Y_6 sont asymptotiquement équivalents.

$$C_1 = a_1^* + \left(\sum_{s_a} \frac{\pi_a}{x_1 x_1'} \right)^{-1} \left(\sum_{s_a} \frac{\pi_a}{x_1 x_2'} \right) a_2^*.$$

avec

$$Y_7 = [X_1' C_1^*] + [M_{x_1} X_2^* a_2^*] + [\hat{Y}^* - \hat{X}_1^* ' a_1^* - \hat{X}_2^* ' a_2^*],$$

on obtient:

En remplaçant X_2^* par son expression trouvée plus haut

$$Y_7 = [X_1' a_1^*] + [X_2^* ' a_2^*] + [\hat{Y}^* - \hat{X}_1^* ' a_1^* - \hat{X}_2^* ' a_2^*].$$

on obtient de la même façon:

En modifiant les poids initiaux en $d_i = F(x_{11} \beta_1) / \pi_a \pi_i$,

$$Y_7 = \hat{Y}^* + [X_1 - \hat{X}_1]' a_1^* + [X_2^* - \hat{X}_2]' a_2^*.$$

l'estimateur associé se réécrit:

$$w_7^i = \frac{F(x_{11} \beta_1) F(x_{11} \delta_1 + x_2' \delta_2)}{\pi_{a1} \pi_i},$$

On obtient des poids

Stratégie 4

équivalentes.

On retrouve donc un estimateur proche de l'estimateur Y_3 proposé dans l'approche modèle de régression dans le cas où l'information contenue dans x_1 est jugée complémentaire à l'information contenue dans x_2 pour estimer y . Le modèle sous-jacent est ici $y_i = x_{11} c_1 + M_{x_1}(x_{12})' c_2 + u_i$. Les différences entre Y_3 et Y_6 portent sur les coefficients estimés: $\left(\frac{\hat{c}_1}{\hat{a}_2} \right)$ diffère légèrement de $\left(\frac{\hat{c}_1}{\hat{a}_2} \right)$ et $[\hat{Y} - \hat{X}_1' a_1 - \hat{X}_2' a_2]$ diffère légèrement de $[\hat{Y} - \hat{X}_1' c_1 - M_{x_1} X_2^* c_2]$. En revanche ces quantités sont asymptotiquement équivalentes.

$$C_1 = a_1 + \left(\sum_{s_a} \frac{\pi_a}{x_1 x_1'} \right)^{-1} \left(\sum_{s_a} \frac{\pi_a}{x_1 x_2'} \right) a_2.$$

avec

$$Y_6 = [X_1' C_1] + [M_{x_1} X_2^* a_2] + [\hat{Y} - \hat{X}_1' a_1 - \hat{X}_2' a_2],$$

On en déduit en remplaçant dans Y_6 que:

$$X_2^* = \sum_{s_a} \frac{\pi_a}{x_2} + \left(\sum_{s_a} \frac{\pi_a}{x_2 x_1'} \right)^{-1} \left(\sum_{s_a} \frac{\pi_a}{x_1 x_1'} \right) \left[X - \sum_{s_a} \frac{\pi_a}{x_1} \right].$$

Or,

Cette stratégie conduit aux équations de calage suivantes :

étape a :

$$\sum_{i \in s_a} \frac{F(x'_{i1} \beta_1)}{\pi_{ai}} x_{i1} = \sum_{i \in U} x_{i1} = X_1$$

qui détermine β_1 , puis

étape b :

$$\sum_{i \in s} \frac{F(x'_{i1} \gamma_1 + x'_{i2} \gamma_2)}{\pi_{ai} \pi_i} x_{i1} = \sum_{i \in U} x_{i1} = X_1, \text{ et}$$

$$\sum_{i \in s_a} \frac{F(x'_{i1} \gamma_1 + x'_{i2} \gamma_2)}{\pi_{ai} \pi_i} x_{i2} = \sum_{i \in s_a} \frac{F(x'_{i1} \beta_1)}{\pi_{ai}} x_{i2} = X_2^*,$$

qui déterminent γ_1 et γ_2 .

Enfin une quatrième stratégie peut être proposée qui peut être vue comme une variante de la stratégie précédente :

Stratégie 4

a) caler la structure de l'échantillon 1^{ère} phase s_a sur celle de la population totale U en terme de variable x_1 , puis,

b) caler la structure de l'échantillon 2^{ème} phase s simultanément en terme de variables x_1 et x_2 , en partant des poids modifiés par le calage précédent, soit :

– sur la structure de la population totale U en ce qui concerne x_1
– sur la structure de s_a modifiée en tenant compte du calage précédent pour x_2 .

Cette stratégie conduit aux équations de calage suivantes :

étape a :

$$\sum_{i \in s_a} \frac{F(x'_{i1} \beta_1)}{\pi_{ai}} x_{i1} = \sum_{i \in U} x_{i1},$$

qui détermine β_1 , puis

étape b :

$$\sum_{i \in s} \frac{F(x'_{i1} \beta_1)}{F(x'_{i1} \beta_1) F(x'_{i2} \delta_1 + x'_{i2} \delta_2)} \pi_{ai} \pi_i x_{i1} = \sum_{i \in U} x_{i1} = X_1, \text{ et}$$

$$\sum_{i \in s_a} \frac{F(x'_{i1} \beta_1)}{F(x'_{i1} \beta_1) F(x'_{i2} \delta_1 + x'_{i2} \delta_2)} \pi_{ai} \pi_i x_{i2} =$$

$$\sum_{i \in s_a} \frac{\pi_{ai}}{F(x'_{i1} \beta_1)} x_{i2} = X_2^*,$$

qui déterminent δ_1 et δ_2 .

Lorsque la fonction de calage est exponentielle, il est clair que les stratégies 3 et 4 coïncident.

Dans cette approche en termes de calage, le point de vue adopté est celui de la réduction de variance basée sur les caractéristiques du plan de sondage, sans considération de modèle. Deux questions se posent alors naturellement :

– Peut-on relier chacune de ces quatre stratégies à une approche en terme de modèle?

– Peut-on comparer ces quatre stratégies en termes de variance?

Nous étudierons d'abord le lien entre les trois stratégies définies par une approche calage et les stratégies définies par une approche modèle ou régression, puis dans un deuxième temps nous intéresserons au calcul des variances des estimateurs associés à chacune des stratégies.

4.2 Lien entre les différentes stratégies possibles et l'approche régression

Lorsque F est linéaire, chacun des estimateurs associés aux quatre stratégies peut se réécrire simplement.

Notations

Dans toute la suite on utilisera les notations suivantes pour un vecteur de variable z quelconque :

$$\hat{z}^* = \sum_{i \in s} \frac{F(x'_{i1} \beta_1)}{\pi_{ai} \pi_i} z_i, \quad \hat{z}^* = \sum_{i \in s_a} \frac{\pi_{ai}}{F(x'_{i1} \beta_1)} z_i.$$

On omettra également les indices i pour alléger la présentation lorsqu'il n'y a pas d'ambiguïté.

Stratégie 1

Les pondérations sont de la forme :

$$w_i^1 = \frac{\pi_{ai} \pi_i}{F(x'_{i1} \beta_1) F(x'_{i2} \beta_2)},$$

l'estimateur associé \hat{Y}_4 peut se réécrire en traduisant l'effet du deuxième calage en x_2 :

$$\hat{Y}_4 = \hat{Y}^* + [\hat{X}_2^* - \hat{X}_2]' \hat{B}_2 \quad \text{avec}$$

$$\hat{B}_2 = \left(\sum_{i \in s} \frac{\pi_{ai} \pi_i}{F(x'_{i1} \beta_1)} x_{i2} x_{i2}' \right)^{-1} \left(\sum_{i \in s} \frac{\pi_{ai} \pi_i}{F(x'_{i1} \beta_1)} x_{i2} y \right),$$

puis en traduisant l'effet du premier calage en x_1 :

$$\hat{Y}_4 = \hat{Y} + [X_1 - \hat{X}_1]' \hat{B}_1 + [\hat{X}_2^* - \hat{X}_2]' \hat{B}_2,$$

soit encore :

$$\hat{Y}_4 = [X_1' \hat{B}_1] + [\hat{X}_2^*]' \hat{B}_2 - \hat{X}_1' \hat{B}_1 + [\hat{Y} - \hat{X}_2^*]' \hat{B}_2,$$

$$\hat{B}_1 = \left(\sum_{i \in s_a} \frac{\pi_{ai}}{x_{i1} x_{i1}'} \right)^{-1} \left(\sum_{i \in s_a} \frac{\pi_{ai}}{x_{i1} y} \right) \quad \text{avec}$$

Avec ces notations les trois estimateurs se réécrivent :

$$Y_1 = [X_1' b_1] + [X_2' b_2 - X_1' b_1] + [\bar{Y} - \hat{X}_2' b_2]$$

associé aux modèles :

$$(1) \quad Y_1 = x_1' b_1 + u_1$$

et

$$(2) \quad Y_1 = x_2' b_2 + u_2$$

$$Y_2 = [X_1' a_1] + [X_2' a_2] + [\bar{Y} - \hat{X}_1' a_1 - \hat{X}_2' a_2]$$

associé au modèle

$$Y_1 = x_1' a_1 + x_2' a_2 + u_1$$

$$Y_3 = [X_1' c_1] + [M_{x_1} X_2' c_2] + [\bar{Y} - \hat{X}_1' c_1 - M_{x_1} \hat{X}_2' c_2]$$

associé au modèle

$$Y_1 = x_1' c_1 + M_{x_1} (x_2)' c_2 + u_1.$$

De la même façon que l'on généralise l'estimateur par la régression par les estimateurs par calage, on peut aborder le problème de l'utilisation de l'information auxiliaire à plusieurs niveaux à travers la théorie du calage, en essayant de construire des stratégies de calage adaptées à la configuration d'information auxiliaire étudiée dans cet article.

4. APPROCHE PAR CALAGE

4.1 Différentes stratégies possibles

Lorsque l'on cherche à généraliser l'estimation par calage proposée dans un contexte où l'information auxiliaire est présente à un seul niveau : celui de la population totale, plusieurs stratégies apparaissent naturellement :

Stratégie 1

a) caler la structure de l'échantillon 1^{ère} phase s_a sur celle de la population totale U en termes de variable x_1 , puis,

b) caler la structure de l'échantillon 2^{ème} phase s sur celle de l'échantillon 1^{ère} phase s_a en terme de variable x_2 .

NB: pour cette dernière opération, mieux vaut tenir compte du calage précédent en x_1 pour déterminer la valeur de référence dans le calage en x_2 sur s_a . Si l'on ne tient pas compte du calage précédent, seules les estimations effectuées au niveau de s_a profiteront de l'amélioration apportée par l'étape a). Une bonne façon de s'en convaincre consiste à considérer le cas particulier extrême où $x_1 = x_2$.

Cette stratégie correspond aux équations de calage suivantes :

Stratégie 3

a) caler la structure de l'échantillon 1^{ère} phase s_a sur celle de la population totale U en termes de variables x_1 , puis,

b) caler la structure de l'échantillon 2^{ème} phase s simultanément en termes de variables x_1 et x_2 , c'est-à-dire :

– sur la structure de la population totale U en ce qui concerne x_1

– sur la structure de s_a modifiée en tenant compte du calage précédent pour x_2 .

pouvoir les dominer :

Une troisième stratégie peut être proposée qui réunit les avantages des deux stratégies précédentes, et semble donc ment d'obtenir une estimation parfaite du total de x_1 :

La première stratégie présente l'avantage de corriger les pondérations 1^{ère} phase, c'est-à-dire d'intégrer l'information auxiliaire au niveau le plus élevé. La deuxième stratégie permet quant à elle, de corriger les pondérations qui seront effectivement utilisées dans l'estimation, et notamment d'obtenir une estimation parfaite du total de x_1 .

qui déterminent α_1 et α_2 .

$$\sum_{i \in s} \frac{F(x_{1i} \alpha_1 + x_{12i} \alpha_2)}{\pi_{ai} \pi_i} x_{1i} = \sum_{i \in U} \frac{F(x_{1i} \alpha_1 + x_{12i} \alpha_2)}{\pi_{ai} \pi_i} x_{1i} = X_1, \text{ et}$$

$$\sum_{i \in s_a} \frac{F(x_{1i} \alpha_1 + x_{12i} \alpha_2)}{\pi_{ai} \pi_i} x_{12i} = \sum_{i \in s} \frac{F(x_{1i} \alpha_1 + x_{12i} \alpha_2)}{\pi_{ai} \pi_i} x_{12i} = X_2,$$

calage suivantes :

Cette deuxième stratégie nous conduit aux équations de

– sur la structure de s_a pour x_2 .

– sur la structure de la population totale U en ce qui concerne x_1

Caler la structure de l'échantillon 2^{ème} phase s simultanément en terme de variables x_1 et x_2 , c'est-à-dire :

Stratégie 2

ou F désigne tout au long de l'article la fonction utilisée dans le calage qui peut être linéaire, exponentielle, linéaire tronquée, logit (cf. Deville, Särndal 1993).

qui détermine β_2 ,

$$\sum_{i \in s} \frac{F(x_{1i} \beta_1)}{\pi_{ai} \pi_i} F(x_{12i} \beta_2) x_{12i} = \sum_{i \in s_a} \frac{F(x_{1i} \beta_1)}{\pi_{ai} \pi_i} F(x_{12i} \beta_2) x_{12i} = X_2^*$$

étape b :

qui détermine β_1 , puis

$$\sum_{i \in s_a} \frac{F(x_{1i} \beta_1)}{\pi_{ai} \pi_i} x_{1i} = \sum_{i \in U} \frac{F(x_{1i} \beta_1)}{\pi_{ai} \pi_i} x_{1i} = X_1$$

étape a :

dont on pense que le second est a priori meilleur pour prédire la valeur de y_i . x_1 fonctionne donc dans cette optique modèle, comme une "proxy" de x_2 . Une situation de ce genre correspond par exemple au cas où x_2 représente la mise à jour c'est-à-dire à la date de l'enquête de la variable extraite de la base de sondage x_1 . Autrement dit si x_2 était disponible au niveau de la population toute entière, l'estimateur utilisé serait

$$\sum_{i \in U} x_{i2}' b_1 + \sum_{i \in s_a} \frac{\pi_i \pi_{ai}}{(y_i - x_{i2}' b_2)}.$$

Imaginons maintenant le cas d'une enquête par sondage en deux phases effectuée auprès des ménages. Supposons que la base de sondage soit constituée de logements pour lesquels on dispose d'une information constituée de la taille du logement notée x_1 , connue donc pour tous les individus de la population cible. Si tous les individus de la première phase de sondage sont interrogés sur la composition du ménage notée x_2 , notamment sur le nombre d'enfants du ménage, les deux informations apparaissent plutôt complémentaires que substituables, lorsqu'il s'agit d'étudier le budget du ménage. Ceci est encore renforcé si au lieu de la composition du ménage on recueille l'âge du chef de ménage ou bien sa profession. Dans une optique modèle, la situation alternative où l'information contenue dans x_1 est jugée complémentaire de celle contenue dans x_2 pour estimer y , apparaît donc naturellement.

3.2 L'information contenue dans x_1 est jugée complémentaire de celle contenue dans x_2 pour estimer y

3.2.1 Décomposition $y_i = x_{i1}' a_1 + x_{i2}' a_2 + u_i$

Le modèle sous-jacent est alors:

$$y_i = x_{i1}' a_1 + x_{i2}' a_2 + u_i \text{ avec } E(u_i) = 0 \text{ et } V(u_i) = \sigma_1^2.$$

L'estimateur à utiliser est alors:

$$Y_2 = \sum_{i \in U} x_{i1}' a_1 + \sum_{i \in s_a} x_{i2}' a_2 + \sum_{i \in U} \frac{\pi_i \pi_{ai}}{(y_i - x_{i1}' a_1 - x_{i2}' a_2)}.$$

avec:

$$\begin{pmatrix} \hat{a}_1 \\ \hat{a}_2 \end{pmatrix} = \begin{pmatrix} \sum_{i \in s_a} \frac{x_{i1}' x_{i1}'}{x_{i1}' x_{i1}'} \pi_{ai} \pi_i & \sum_{i \in s_a} \frac{x_{i1}' x_{i2}'}{x_{i1}' x_{i1}'} \pi_{ai} \pi_i \\ \sum_{i \in s_a} \frac{x_{i2}' x_{i1}'}{x_{i1}' x_{i1}'} \pi_{ai} \pi_i & \sum_{i \in s_a} \frac{x_{i2}' x_{i2}'}{x_{i1}' x_{i1}'} \pi_{ai} \pi_i \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i \in s_a} \frac{x_{i1}' y_i}{x_{i1}' x_{i1}'} \pi_{ai} \pi_i \\ \sum_{i \in s_a} \frac{x_{i2}' y_i}{x_{i1}' x_{i1}'} \pi_{ai} \pi_i \end{pmatrix}.$$

En effet, la variable y se décompose en trois composantes $y_i = x_{i1}' a_1 + x_{i2}' a_2 + u_i$. Le total de y se décompose

donc en trois composantes qui sont chacune estimée au niveau le plus haut, c'est-à-dire avec le maximum de

précision possible:

- U pour $x_{i1}' a_1$,
- s_a pour $x_{i2}' a_2$, et
- s pour u_i .

3.2.2 Décomposition $y_i = x_{i1}' c_1 + M_{x_1}(x_{i2})' c_2 + u_i$

Si on souhaite utiliser au maximum l'information contenue dans x_1 disponible sur U , il est naturel d'introduire une autre écriture du même modèle $y_i = x_{i1}' a_1 + x_{i2}' a_2 + u_i$ qui isole tout ce qui dans y peut être pris en compte à travers x_1 . Elle s'écrit:

$$y_i = x_{i1}' c_1 + M_{x_1}(x_{i2})' c_2 + u_i \text{ avec}$$

$$E(u_i) = 0 \text{ et } V(u_i) = \sigma_1^2,$$

où M_{x_1} représente la projection orthogonale, dans la métrique associée aux poids $1/\pi_{ai}$, sur l'orthogonal de l'espace vectoriel engendré dans s_a (assimilé à \mathcal{H}^n) par le groupe de variables x_1 .

$M_{x_1}(x_{i2})$ est définie par:

$$M_{x_1}(x_{i2}) = x_{i2} - \left(\sum_{i \in s_a} \frac{x_{i2}' x_{i1}'}{x_{i1}' x_{i1}'} \pi_{ai} \right) \left(\sum_{i \in s_a} \frac{x_{i1}' x_{i1}'}{x_{i1}' x_{i1}'} \pi_{ai} \right)^{-1} x_{i1}.$$

L'estimateur naturel associé est alors:

$$Y_3 = \sum_{i \in U} x_{i1}' c_1 + \sum_{i \in s_a} \frac{\pi_{ai}}{(M_{x_1} x_{i2})' c_2)} + \sum_{i \in s_a} \frac{\pi_i \pi_{ai}}{(y_i - x_{i1}' c_1 - M_{x_1} x_{i2}' c_2)},$$

où $\hat{c} = \begin{pmatrix} \hat{c}_1 \\ \hat{c}_2 \end{pmatrix}$ est le coefficient de la régression $y = x' c_1 + (M_{x_1} x_2)' c_2 + u$ estimée sur s avec des poids $1/\pi_{ai} \pi_i$ (qui diffère légèrement de $\begin{pmatrix} \hat{a}_1 \\ \hat{a}_2 \end{pmatrix}$).

3.3 Les trois estimateurs dérivés de l'approche modèles

L'approche par la modélisation nous a permis de constater 3 estimateurs que l'on peut récrire synthétiquement en introduisant de nouvelles notations. On écrira dans toute la suite pour un vecteur de variable z quelconque:

$$\hat{Z} = \sum_{i \in s_a} \frac{\pi_{ai} \pi_i}{1} z, \quad \hat{\tilde{Z}} = \sum_{i \in s_a} \frac{\pi_{ai} \pi_i}{1} z.$$

Ainsi (section 4), les deux approches peuvent être reliées et débouchent sur trois classes d'estimateurs associées chacune à une décomposition de la variable d'intérêt. Les estimateurs d'une même classe ont la même variance asymptotique.

L'évaluation des stratégies fondée sur le seul plan de sondage oriente le choix vers la troisième classe d'estimateurs qui domine les deux autres du point de vue de la variance.

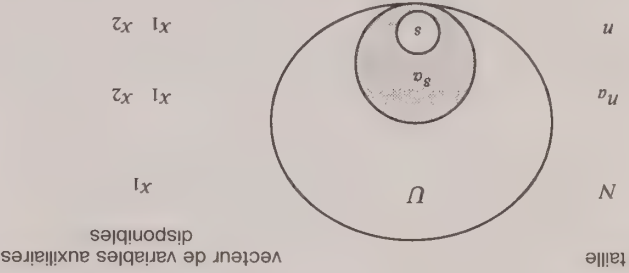
L'évaluation des stratégies, lorsqu'elle s'appuie sur une modélisation de la variable d'intérêt conduit à privilégier la classe d'estimateurs associée à la modélisation adoptée. Dans un contexte où l'on souhaite effectuer le redressement d'une enquête, et que l'on souhaite simultanément corriger les biais qu'induirait l'utilisation des pondérations brutes et réduire la variance, les conclusions doivent être adaptées: les modifications introduites dans les pondérations pour corriger les biais sont plus importantes que les corrections pour réduction de variance. Dès lors, on intégrera dans le calage les variables, dès que l'on pense qu'elles interviennent dans la probabilité de sélection et donc participent à la création du biais.

Lorsque les variables auxiliaires sont qualitatives, l'arbitrage entre utilisation a priori et a posteriori de l'information auxiliaire, c'est-à-dire entre, utilisation au niveau de l'échantillonnage ou au niveau du redressement, repose encore sur la distinction entre les deux modélisations de la variable d'intérêt.

Il est possible d'étendre ces résultats au cas de sondages en plus de deux phases.

2. NOTATIONS

Le cadre est celui d'un sondage en deux phases. On suppose qu'on dispose d'information auxiliaire à deux niveaux différents: population cible et population issue du tirage 1^{ère} phase. La situation peut être schématisée de la manière suivante:



Où U représente la population cible pour laquelle les valeurs du vecteur de variable x_1 sont connues ou à défaut le total $X_1 = \sum_{i \in U} x_{i1}$. s_a représente un niveau intermédiaire de sondage pour lequel les valeurs des vecteurs de k_1 variables x_1 et k_2 variables x_2 sont connues pour tous les individus. On note π_{ia} , la probabilité de sélection du tirage associée à la première phase du sondage. s représente l'échantillon final pour lequel les valeurs de la variable y

3.1 L'information contenue dans x_1 est jugée substituable à l'information contenue dans x_2 pour estimer y et de moins bonne qualité que x_2
Särndal, Swensson et Wretman proposent dans leur ouvrage l'estimateur par régression suivant pour l'estimation du total de y :

$$Y_1 = \sum_{i \in s} \frac{y_i \pi_i \pi_{ai}}{x'_{i2} b_2} + \left(\sum_{i \in s_a} \frac{\pi_i}{x'_{i2} b_2} - \sum_{i \in s} \frac{\pi_i \pi_{ai}}{x'_{i2} b_2} \right) + \left(\sum_{i \in U} x'_{i1} b_1 - \sum_{i \in s_a} \frac{\pi_{ai}}{x'_{i1} b_1} \right)$$

où le deuxième terme est la correction pour mauvaise estimation sur s_a et le troisième est la correction pour mauvaise estimation sur s .

L'estimateur peut aussi s'écrire:

$$Y_1 = \sum_{i \in U} x'_{i1} b_1 + \sum_{i \in s_a} \frac{\pi_{ai}}{(x'_{i1} b_1 - x'_{i2} b_2)} + \sum_{i \in s} \frac{\pi_i \pi_{ai}}{(y_i - x'_{i2} b_2)}$$

où le deuxième terme est la correction pour mauvaise approximation de y_i par $x'_{i1} b_1$ et le troisième est la correction pour mauvaise approximation de y_i par $x'_{i2} b_2$;

$$b_1 = \left(\sum_{i \in s_a} \frac{\pi_{ai}}{x'_{i1} x'_{i1}} \right)^{-1} \left(\sum_{i \in s_a} \frac{\pi_{ai}}{x'_{i1} y_i} \right) \quad \text{avec}$$

$$b_2 = \left(\sum_{i \in s} \frac{\pi_i \pi_{ai}}{x'_{i2} x'_{i2}} \right)^{-1} \left(\sum_{i \in s} \frac{\pi_i \pi_{ai}}{x'_{i2} y_i} \right) \quad \text{et}$$

L'idée sous-jacente est que l'on a deux modèles concurrents pour y qui sont:

$$(1) y_i = x'_{i1} b_1 + u_{i1} \quad \text{avec} \quad E(u_{i1}) = 0 \quad \text{et} \quad V(u_{i1}) = \sigma_1^2$$

$$(2) y_i = x'_{i2} b_2 + u_{i2} \quad \text{avec} \quad E(u_{i2}) = 0 \quad \text{et} \quad V(u_{i2}) = \sigma_2^2$$

Redressements alternatifs en présence de plusieurs niveaux d'information auxiliaire

F. DUPONT¹

RÉSUMÉ

L'estimation par régression et sa généralisation, l'estimation par calage introduite par Deville et Särndal en 1993 permet de réduire la variance des estimateur, grâce à l'utilisation de l'information auxiliaire. Dans les enquêtes réalisées par sondage, on dispose souvent d'information additionnelle exploitable répartie selon un schéma complexe notamment lorsque le sondage est réalisé en plusieurs phases. Une adaptation de l'estimation par régression a été proposée avec ses variantes dans le cadre du sondage à deux phases par Särndal et Swensson en 1987. Cet article propose d'étudier les stratégies d'estimation alternatives selon deux configurations alternatives pour l'information auxiliaire, en reliant les deux approches possibles pour aborder le problème: modèle de régression et estimation par calage.

MOTS CLÉS: Information auxiliaire; estimateur par régression; estimateur par calage; sondage en deux phases.

1. INTRODUCTION

L'estimateur par régression étudié par Fuller (1975), Cassel, Särndal et Wretman (1976), Särndal (1980), Goutroux (1981), Isaki et Fuller (1982), et Wright (1983) permet d'améliorer, a posteriori, c'est-à-dire une fois le sondage réalisé, l'estimation d'un total d'une variable d'intérêt sur la base de variables auxiliaires x_1, \dots, x_k pour lesquelles on dispose d'information additionnelle. On réduit en effet la variance par rapport à l'estimateur d'Horvitz-Thompson en utilisant l'estimateur par régression à condition de connaître la valeur sans aléa, des totaux sur la population cible des variables auxiliaires, qui constituent l'information additionnelle appelée information auxiliaire. Deville et Särndal ont proposé en 1992 une classe d'estimateurs dérivée d'une approche par pondération qui répond à la même problématique de réduction de variance: les estimateurs par calage. Le calage des poids de sondage réalisé permet en effet d'intégrer a posteriori une information auxiliaire du type totaux X_1, \dots, X_k de k variables x_1, \dots, x_k dans l'estimation réalisée à partir des nouvelles pondérations et donc de l'améliorer. Cette approche généralise l'estimation par régression qui constitue l'un des éléments de la classe.

Toutefois, dans les enquêtes réalisées par sondage, on dispose souvent d'information additionnelle exploitable répartie selon un schéma plus complexe que ce qui vient d'être décrit, notamment lorsque le sondage est réalisé en plusieurs phases. Cet article étudie différentes stratégies d'utilisation de cette information auxiliaire complexe dans le cadre d'un sondage en deux phases, la généralisation à plus de deux phases étant possible.

Lorsque le plan de sondage comporte deux phases, l'information auxiliaire est constituée de l'information connue pour la population toute entière, mais aussi de l'information connue pour l'échantillon issu de la première

phase de sondage. Or, ces deux informations peuvent porter sur des variables différentes.

Särndal et Swensson proposent dans leur article de 1987, un estimateur utilisant toute l'information auxiliaire disponible pour un tirage en deux phases, avec une information auxiliaire différente pour la population toute entière et la population issu du tirage de la première phase. Il s'agit d'un estimateur adaptant le principe de l'estimateur par régression lorsque l'information connue pour les individus issus du tirage première phase est considérée comme substituable à l'information agrégée et de meilleure qualité que l'information disponible pour la population cible toute entière, vis-à-vis de l'estimation de la variable d'intérêt. Or, il arrive en pratique que ces deux informations soient complémentaires plutôt que substituables. On a donc cherché dans cette étude à développer l'estimation par régression, dans un contexte où les informations auxiliaires sont complémentaires.

Par ailleurs, dans la mesure où l'estimation par calage généralise l'estimation par régression lorsque l'information auxiliaire n'est constituée que d'un seul niveau, on a cherché à adapter l'estimation par calage dans ce contexte. On passe en revue les stratégies de calages afin de proposer les plus adaptées, en cherchant à les relier aux généralisations de l'estimation par régression possibles dans ce contexte.

On montre (section 2) que l'utilisation conjointe de deux informations auxiliaires différentes conduit à deux modèles de régression et trois décompositions de la variable d'intérêt associées. L'approche assistée par modèle de régression (AAMR) permet donc de dériver 3 estimateurs alternatifs.

L'approche calage (AC) (section 3) permet quant à elle de dériver 4 estimateurs. Chacun de ces estimateurs peut être relié (associé) aux trois estimateurs dérivés de l'approche modèle de régression.

¹ F. Dupont, Unité Méthodes Statistiques, Institut National de la Statistique et des Etudes Economiques, (INSEE), 18 Blvd. Adolphe Pinard, 75675 Paris Cedex.

LE, T., et VERMA, V. (1995). *Sample Designs and Sampling Errors for the DHS*. Calverton MD: MACRO International. John Wiley and Sons.

McNEMAR, Q. (1949). *Psychological Statistics*. New York: John Wiley and Sons.

MOSTELLER, F. (1952). Some statistical problems in measuring the subjective responses to drugs. *Biometrika*, 8, 220-226.

RAO, J.N.K., et SCOTT, A.J. (1987). On simple adjustments to Chi-square tests with sample survey data. *Annals of Statistics*, 15, 385-397.

RAO, J.N.K., et WU, C.F.S. (1985). Inference from stratified samples: Second-order analysis of three methods for nonlinear statistics. *Journal of the American Statistical Association*, 80, 620-630.

SCOTT, A.J., et HOLT, D. (1982) The effect of two-stage sampling on ordinary least squares methods. *Journal of the American Statistical Association*, 77, 848-54.

SCOTT, A.J., et SEBER, G.A.F. (1983). Difference of proportions from the same survey. *The American Statistician*, 37, 319-20.

SKINNER, C.J., HOLT, D., et SMITH, T.M.F. (1989). *Analysis of Complex Surveys*. New York: John Wiley and Sons.

VERMA, V., et LE, T. (1995). Sampling errors for the DHS survey. 50ième Session de l'Institut International de Statistique, Beijing.

VERMA, V., SCOTT, C., et O'MURCHARTAIGH, C. (1980). Sample designs and sampling errors for the World Fertility Surveys. *Journal of the Royal Statistical Society (A)*, 143, 431-473.

WILD, C.J., et SEBER, G.A.F. (1993). Comparing two proportions for the same survey. *The American Statistician*, 47, 178-181.

tableaux 1, 2 et 3. La figure 1 présente les données provenant de trois pays (Maroc, Niger et Colombie); il y a donc six populations puisque les eps en zone urbaine sont assez différents de ceux observés en zone rurale. La figure 2 présente les résultats selon le sexe; on distingue deux populations passablement distinctes en ce qui concerne l'emploi, mais la différence est moins marquée en ce qui concerne le niveau d'éducation.

Nous attendions les données empiriques des tableaux des études 4, 5 et 6 avec anxiété. Il est vrai que les cinq ensembles précédents avaient conduit à des conclusions similaires, même s'ils traitaient de onze populations différentes, de pointages et de variables. Pourtant, les études 1 à 5 portaient sur des paires de catégories issues de polytomies, les plans 1 et 2 du type A. D'ores et déjà, nous cherchions des données pour des comparaisons de type B provenant d'enquêtes par panel, de manière à pouvoir explorer les conjectures des plans test-retest et avant-après. Du point de vue mathématique, on peut facilement montrer une ressemblance avec les polytomies (c.-à-d., avec les tétratomies), mais les rapports entre ces valeurs et les valeurs empiriques des eps ne sont pas du tout évidentes. C'est ce qui explique pourquoi ces valeurs empiriques sont si utiles et si remarquables. Nous avons ici observé un effet du plan de sondage extrêmement important pour les tests de chi carré des comparaisons analytiques.

5. SIGNIFICATION DES RÉSULTATS POUR LES RECHERCHES CONNEXES

Nos travaux antérieurs et ceux d'autres chercheurs fournissent déjà une masse considérable de données empiriques sur les effets du plan de sondage sur l'échantillon total, les sous-classes et les différences, pour des variables et des plans différents. Il serait donc utile d'examiner les résultats décrits plus haut à la lumière de ces données antérieures.

On a déterminé que la nature des variables d'enquête qui font l'objet d'une estimation est un facteur important (souvent le plus important) pour déterminer l'ampleur des eps. On peut observer des eps extrêmement différents selon le type de variables, même au sein d'un seul échantillon ou avec des plans de sondage très semblables. C'est la raison pour laquelle nous avons toujours recommandé que les eps soient calculés pour plusieurs variables différentes, alors qu'il est généralement moins important de les calculer pour plusieurs sous-classes différentes, en particulier lorsque ces sous-classes sont définies en fonction des mêmes caractéristiques.

Nos résultats montrent que les eps peuvent également varier énormément entre les catégories différentes de la même variable d'enquête, lorsque l'échantillon total sert de base commune à l'estimation. Ainsi, il convient dans un certain sens de considérer chaque catégorie individuelle et chaque différence entre les paires de catégories, même lorsqu'on a affaire aux variables d'une seule et même enquête, comme s'il s'agissait de variables séparées aux fins du calcul et de l'analyse des effets du plan de sondage.

Pour ce qui est du rapport existant entre les eps pour les sous-classes et pour les différences entre sous-classes, les recherches antérieures ont surtout porté sur la situation décrite ci-après. Compte tenu d'un échantillon total n réparti en sous-classes i de taille $n_i = p_i \cdot n$, les valeurs $d_{\text{eps}}(r_i)$ pour les statistiques r_i (comme le rapport m_i/n_i , la moyenne $\sum y_i/n_i$ ou le rapport $\sum y_i/\sum x_i$), calculées sur la base des éléments de sous-classe n_i , se rapportent à la valeur $\text{eps}(r)$ pour la même variable, calculée sur la base de l'échantillon total. De même, les valeurs de $\text{eps}(r_i - r_j)$ pour les différences de sous-classes se rapportent à l'eps(r_i), à l'eps(r_j) fondé sur les sous-classes individuelles et à l'eps(r) fondé sur l'échantillon total. De nombreux calculs confirment que ces rapports s'accordent avec notre conjecture (2) de la section 3:

$$\text{eps}(r) > \text{eps}(r_i); \text{ et } \text{eps}(r_i) > \text{eps}(r_i - r_j) > 1.$$

Ces effets des covariances sur les effets du plan de sondage pour des échantillons par grappes sont essentiellement empiriques (même sociologiques, au sens large); ils doivent être vérifiés comme tels.

Ainsi en est-il du nouveau rapport que nous avons découvert pour $(p_i - p_j)$ pour deux catégories, lesquelles sont si différentes de ce qui précède. Les rapports $\text{eps}(p_i - p_j) \equiv [\text{eps}(p_i) + \text{eps}(p_j)]/2$ sont également empiriques et approximatifs, et doivent être constamment vérifiés. Pourtant, ils semblent s'appliquer largement à nos données et sont nettement préférables aux autres hypothèses, telles que $\text{eps}(p_i - p_j) = 1$, qu'on a souvent utilisées jusqu'ici.

REMERCIEMENTS

Les auteurs désirent remercier l'éditeur associé et les arbitres dont les commentaires ont permis d'abréger et d'améliorer cet article.

BIBLIOGRAPHIE

- COCHRAN, W.G. (1950). The comparison of percentages in matched samples. *Biometrika*, 37, 256-66.
- DEMING, W.E. (1953). On the distinction between enumerative and analytic studies. *Journal of the American Statistical Association*, 48, 244-45.
- KISH, L. (1965). *Survey Sampling*. New York: John Wiley and Sons.
- KISH, L. (1987). *Statistical Research Design*. New York: John Wiley and Sons.
- KISH, L. (1995). Methods for design effects. *Journal of Official Statistics*, 11, 55-77.
- KISH, L., et FRANKEL, M.R. (1974). Inference from complex samples. *Journal of the Royal Statistical Society*, (B), 36, 1-37.
- KISH, L., GROVES, R.M., et KROTKE, K. (1976). *Sampling Errors for Fertility Surveys*. Document hors série n° 17, *Enquête mondiale de la fécondité*. Institut International de Statistique: The Hague.

des quelques cas où $\text{eps}(p_i - p_j)$ ne se situe pas entre $\text{eps}(p_i)$ et $\text{eps}(p_j)$, et où on obtient $\text{eps}(p_i) < \text{eps}(p_i - p_j) > \text{eps}(p_j)$ ou $\text{eps}(p_i) > \text{eps}(p_i - p_j) < \text{eps}(p_j)$. Ces cas montrent en passant que nos résultats ne sont pas des conséquences mathématiques, mais qu'ils ont plutôt un caractère empirique.

Tableau 4

Les National Election Studies Panels de 1990 et de 1992, réalisés par le Survey Research Center, Institute for Social Research, Ann Arbor

Catégories	P_i	P_j	Moyenne	avant/après (90/92)
Eps pour	$(P_i - P_j)$			

Entièrement d'accord avec Bush	1.14	.93	1.04	1.02
D'accord avec la politique étrangère de Bush	.92	1.05	.99	1.00
Désapprouve entièrement la politique étrangère de Bush	1.23	1.24	1.24	1.32
D'accord avec la politique économique de Bush	.97	.94	.96	.96
Entièrement d'accord avec la politique économique de Bush	1.14	1.04	1.09	1.10
D'accord avec Bush	1.00	1.00	1.00	1.00
Désapprouve entièrement Bush	1.16	1.10	1.13	1.12
A suivi la campagne électorale à la télé	.89	1.55	1.22	1.40
Moyenne	1.06	1.11	1.08	1.11

Tableau 6

Les études Americans' Changing Lives de 1986 et de 1989 réalisées par le Survey Research Center, Ann Arbor

Catégories	P_i	P_j	Moyenne	avant/après
Eps pour	$(P_i - P_j)$			

A. Rencontre des amis	1.30	1.26	1.28	1.28
Une fois par semaine	.88	1.00	.94	1.02
2 ou 3 fois par mois	1.09	1.13	1.11	1.15
Moyenne	1.28	1.21	1.25	1.33
Très satisfait	1.04	1.16	1.10	1.00
Pas satisfait	1.16	1.19	1.18	1.17
E. Pratique le jardinage	1.40	1.16	1.28	1.19
Souvent	.91	1.11	1.01	1.18
Rarement	1.66	1.17	1.42	1.26
Jamais	1.32	1.15	1.24	1.21
Moyenne	1.25	1.26	1.26	1.18
Moyenne globale	1.25	1.26	1.26	1.18
B. Fait de l'exercice physique	1.51	1.67	1.59	1.26
Souvent	1.62	1.97	1.80	1.41
Jamais	1.56	1.82	1.70	1.34
Moyenne	1.24	.90	1.07	.91
Beaucoup	1.33	.98	1.16	1.12
Pas beaucoup	1.29	.94	1.12	1.02
Moyenne	1.08	1.31	1.20	1.20
F. A une attitude positive	1.10	1.33	1.22	1.19
Oui	1.05	1.28	1.17	1.21
Non	1.08	1.31	1.20	1.20
Moyenne	1.19	1.22	1.20	1.19
D. Aime sa maison	1.51	1.67	1.59	1.26
Souvent	1.62	1.97	1.80	1.41
Jamais	1.56	1.82	1.70	1.34
Moyenne	1.24	.90	1.07	.91
Beaucoup	1.33	.98	1.16	1.12
Pas beaucoup	1.29	.94	1.12	1.02
Moyenne	1.08	1.31	1.20	1.20
F. A une attitude positive	1.10	1.33	1.22	1.19
Oui	1.05	1.28	1.17	1.21
Non	1.08	1.31	1.20	1.20
Moyenne	1.19	1.22	1.20	1.19

Tableau 5

Les Panel Study of Income Dynamics de 1983 et de 1987, réalisés par le Survey Research Center, Ann Arbor

Catégories*	P_i	P_j	Moyenne	avant/après (83/87)
Eps pour	$(P_i - P_j)$			

Vit dans le sud	1.22	1.23	1.23	1.11
Âge du chef de famille	1.28	1.33	1.31	1.37
Taille de la famille	1.29	1.43	1.36	1.47
Nombre d'enfants dans la famille	1.23	1.43	1.33	1.49
Heures de travail du chef de famille	1.12	.84	.98	1.03
Âge de l'enfant le plus jeune	.93	.91	.92	.87
Moyenne	1.18	1.20	1.19	1.22

* Toutes les variables sont classées dans deux catégories.

Les résultats empiriques présentés aux figures 1 et 2 viennent encore confirmer ceux présentés aux tableaux 1, 2 et 3. Nous y constatons également que: 1) $\text{eps}(p_i - p_j) \equiv [\text{eps}(p_i) + \text{eps}(p_j)] / 2$ approximativement, le long de la ligne de 45°; 2) ces égalités se vérifient pour une vaste gamme d'eps; 3) la variation entre les variables est en effet très importante. Cette variation est particulièrement évidente pour l'Indonésie rurale, où les valeurs d'eps dépassent 4 et où les valeurs d'eps² sont donc supérieures à 16. Ces importants effets de groupement sont dus à la grande taille des grappes: avec $b = 133$ et 137, les valeurs de $roh = 0.12$ sont suffisantes pour un grand eps. À noter que ces résultats empiriques viennent à la fois de populations et de variables très diversifiées; elles sont différentes l'une de l'autre et différentes également des données des

Tableau 2
La National Education Longitudinal Study (NELS) de 1988, réalisée par le National Opinion Research Center de la University of Chicago ($n = 24,355$)

Catégories		$i - j$	P_i	P_j	Moyenne	$(P_i - P_j)$	Eps pour	
A. Scolarité								
1-2	1.38	1.22	1.30	1.11	1.26	1.16	1-2	1.04
1-3	1.38	1.14	1.26	1.16	1.30	1.14	1-3	1.07
1-4	1.38	1.19	1.29	1.16	1.27	1.12	1-4	1.13
1-5	1.38	1.42	1.40	1.54	1.30	1.08	1-5	1.08
2-3	1.22	1.14	1.18	1.11	1.24	1.13	2-3	1.11
2-4	1.22	1.19	1.21	1.24	1.24	1.11	2-4	1.13
2-5	1.22	1.42	1.32	1.45	1.24	1.08	2-5	1.14
3-4	1.14	1.19	1.17	1.18	1.28	1.14	3-4	1.14
3-5	1.14	1.42	1.28	1.37	1.20	1.08	3-5	1.15
4-5	1.19	1.42	1.31	1.20	1.25	1.04	4-5	1.17
E. Religion								
1-2	2.48	2.83	2.65	2.74	2.09	1.24	1-2	1.17
1-3	2.48	2.02	2.25	2.09	1.18	1.16	1-3	1.24
1-4	2.48	2.02	2.42	2.59	1.18	1.16	1-4	1.24
1-5	2.48	2.60	2.29	2.47	1.18	1.16	1-5	1.24
2-3	2.83	2.02	2.42	2.59	1.18	1.16	2-3	1.14
2-4	2.83	2.02	2.42	2.59	1.18	1.16	2-4	1.14
2-5	2.83	2.02	2.42	2.59	1.18	1.16	2-5	1.14
3-4	1.14	1.19	1.17	1.18	1.28	1.14	3-4	1.14
3-5	1.14	1.42	1.28	1.37	1.20	1.08	3-5	1.15
4-5	1.19	1.42	1.31	1.20	1.25	1.04	4-5	1.17
I. Assurance								
1-2	1.42	1.28	1.35	1.37	1.65	1.83	1-2	1.83
1-3	1.42	1.68	1.61	1.65	1.72	1.65	1-3	1.65
1-4	1.42	1.76	1.61	1.65	1.72	1.65	1-4	1.65
1-5	1.42	1.76	1.61	1.65	1.72	1.65	1-5	1.65
2-3	1.68	1.68	1.76	1.65	1.72	1.65	2-3	1.65
2-4	1.68	1.68	1.76	1.65	1.72	1.65	2-4	1.65
2-5	1.68	1.68	1.76	1.65	1.72	1.65	2-5	1.65
3-4	1.68	1.68	1.76	1.65	1.72	1.65	3-4	1.65
3-5	1.68	1.68	1.76	1.65	1.72	1.65	3-5	1.65
4-5	1.68	1.68	1.76	1.65	1.72	1.65	4-5	1.65
F. Scolarité du père								
1-2	1.61	1.76	1.61	1.65	1.72	1.65	1-2	1.65
1-3	1.61	1.68	1.61	1.65	1.72	1.65	1-3	1.65
1-4	1.61	1.68	1.61	1.65	1.72	1.65	1-4	1.65
1-5	1.61	1.68	1.61	1.65	1.72	1.65	1-5	1.65
2-3	1.76	1.68	1.76	1.65	1.72	1.65	2-3	1.65
2-4	1.76	1.68	1.76	1.65	1.72	1.65	2-4	1.65
2-5	1.76	1.68	1.76	1.65	1.72	1.65	2-5	1.65
3-4	1.76	1.68	1.76	1.65	1.72	1.65	3-4	1.65
3-5	1.76	1.68	1.76	1.65	1.72	1.65	3-5	1.65
4-5	1.76	1.68	1.76	1.65	1.72	1.65	4-5	1.65
D. L'école donne accès à de bons emplois								
1-2	1.24	1.07	1.24	1.07	1.16	1.04	1-2	1.04
1-3	1.24	1.11	1.24	1.11	1.16	1.04	1-3	1.04
1-4	1.24	1.11	1.24	1.11	1.16	1.04	1-4	1.04
1-5	1.24	1.11	1.24	1.11	1.16	1.04	1-5	1.04
2-3	1.07	1.11	1.07	1.11	1.16	1.04	2-3	1.04
2-4	1.07	1.11	1.07	1.11	1.16	1.04	2-4	1.04
2-5	1.07	1.11	1.07	1.11	1.16	1.04	2-5	1.04
3-4	1.07	1.11	1.07	1.11	1.16	1.04	3-4	1.04
3-5	1.07	1.11	1.07	1.11	1.16	1.04	3-5	1.04
4-5	1.07	1.11	1.07	1.11	1.16	1.04	4-5	1.04
C. Libre de choisir sa voie								
1-2	1.28	1.10	1.28	1.10	1.19	1.21	1-2	1.21
1-3	1.28	1.08	1.28	1.08	1.19	1.21	1-3	1.21
1-4	1.28	1.08	1.28	1.08	1.19	1.21	1-4	1.21
1-5	1.28	1.08	1.28	1.08	1.19	1.21	1-5	1.21
2-3	1.10	1.08	1.10	1.08	1.19	1.21	2-3	1.08
2-4	1.10	1.08	1.10	1.08	1.19	1.21	2-4	1.08
2-5	1.10	1.08	1.10	1.08	1.19	1.21	2-5	1.08
3-4	1.10	1.08	1.10	1.08	1.19	1.21	3-4	1.08
3-5	1.10	1.08	1.10	1.08	1.19	1.21	3-5	1.08
4-5	1.10	1.08	1.10	1.08	1.19	1.21	4-5	1.08
B. Trouve les cours ennuyants								
1-2	.99	1.11	.99	1.05	1.07	1.04	1-2	1.04
1-3	.99	1.12	.99	1.06	1.07	1.04	1-3	1.07
1-4	.99	1.12	.99	1.06	1.07	1.04	1-4	1.07
1-5	.99	1.12	.99	1.06	1.07	1.04	1-5	1.07
2-3	1.11	1.12	1.11	1.12	1.07	1.04	2-3	1.07
2-4	1.11	1.12	1.11	1.12	1.07	1.04	2-4	1.07
2-5	1.11	1.12	1.11	1.12	1.07	1.04	2-5	1.07
3-4	1.11	1.12	1.11	1.12	1.07	1.04	3-4	1.07
3-5	1.11	1.12	1.11	1.12	1.07	1.04	3-5	1.07
4-5	1.11	1.12	1.11	1.12	1.07	1.04	4-5	1.07
D. L'école donne accès à de bons emplois								
1-2	1.24	1.07	1.24	1.07	1.16	1.04	1-2	1.04
1-3	1.24	1.11	1.24	1.11	1.16	1.04	1-3	1.04
1-4	1.24	1.11	1.24	1.11	1.16	1.04	1-4	1.04
1-5	1.24	1.11	1.24	1.11	1.16	1.04	1-5	1.04
2-3	1.07	1.11	1.07	1.11	1.16	1.04	2-3	1.04
2-4	1.07	1.11	1.07	1.11	1.16	1.04	2-4	1.04
2-5	1.07	1.11	1.07	1.11	1.16	1.04	2-5	1.04
3-4	1.07	1.11	1.07	1.11	1.16	1.04	3-4	1.04
3-5	1.07	1.11	1.07	1.11	1.16	1.04	3-5	1.04
4-5	1.07	1.11	1.07	1.11	1.16	1.04	4-5	1.04
C. Libre de choisir sa voie								
1-2	1.28	1.10	1.28	1.10	1.19	1.21	1-2	1.21
1-3	1.28	1.08	1.28	1.08	1.19	1.21	1-3	1.21
1-4	1.28	1.08	1.28	1.08	1.19	1.21	1-4	1.21
1-5	1.28	1.08	1.28	1.08	1.19	1.21	1-5	1.21
2-3	1.10	1.08	1.10	1.08	1.19	1.21	2-3	1.08
2-4	1.10	1.08	1.10	1.08	1.19	1.21	2-4	1.08
2-5	1.10	1.08	1.10	1.08	1.19	1.21	2-5	1.08
3-4	1.10	1.08	1.10	1.08	1.19	1.21	3-4	1.08
3-5	1.10	1.08	1.10	1.08	1.19	1.21	3-5	1.08
4-5	1.10	1.08	1.10	1.08	1.19	1.21	4-5	1.08
F. Scolarité du père								
1-2	1.61	1.76	1.61	1.65	1.72	1.65	1-2	1.65
1-3	1.61	1.68	1.61	1.65	1.72	1.65	1-3	1.65
1-4	1.61	1.68	1.61	1.65	1.72	1.65	1-4	1.65
1-5	1.61	1.68	1.61	1.65	1.72	1.65	1-5	1.65
2-3	1.76	1.68	1.76	1.65	1.72	1.65	2-3	1.65
2-4	1.76	1.68	1.76	1.65	1.72	1.65	2-4	1.65
2-5	1.76	1.68	1.76	1.65	1.72	1.65	2-5	1.65
3-4	1.76	1.68	1.76	1.65	1.72	1.65	3-4	1.65
3-5	1.76	1.68	1.76	1.65	1.72	1.65	3-5	1.65
4-5	1.76	1.68	1.76	1.65	1.72	1.65	4-5	1.65
E. Scolarité								
1-2	1.42	1.28	1.35	1.37	1.65	1.83	1-2	1.83
1-3	1.42	1.68	1.61	1.65	1.72	1.65	1-3	1.65
1-4	1.42	1.76	1.61	1.65	1.72	1.65	1-4	1.65
1-5	1.42	1.76	1.61	1.65	1.72	1.65	1-5	1.65
2-3	1.68	1.68	1.76	1.65	1.72	1.65	2-3	1.65
2-4	1.68	1.68	1.76	1.65	1.72	1.65	2-4	1.65
2-5	1.68	1.68	1.76	1.65	1.72	1.65	2-5	1.65
3-4	1.68	1.68	1.76	1.65	1.72	1.65	3-4	1.65
3-5	1.68	1.68	1.76	1.65	1.72	1.65	3-5	1.65
4-5	1.68	1.68	1.76	1.65	1.72	1.65	4-5	1.65
I. Assurance								
1-2	1.42	1.28	1.35	1.37	1.65	1.83	1-2	1.83
1-3	1.42	1.68	1.61	1.65	1.72	1.65	1-3	1.65
1-4	1.42	1.76	1.61	1.65	1.72	1.65	1-4	1.65
1-5	1.42	1.76	1.61	1.65	1.72	1.65	1-5	1.65
2-3	1.68	1.68	1.76	1.65	1.72	1.65	2-3	1.65
2-4	1.68	1.68	1.76	1.65	1.72	1.65	2-4	1.65
2-5	1.68	1.68	1.76	1.65	1.72	1.65	2-5	1.65
3-4	1.68	1.68	1.76	1.65	1.72	1.65	3-4	1.65
3-5	1.68	1.68	1.76	1.65	1.72	1.65	3-5	1.65
4-5	1.68	1.68	1.76	1.65	1.72	1.65	4-5	1.65
D. L'école donne accès à de bons emplois								
1-2	1.24	1.07	1.24	1.07	1.16	1.04	1-2	1.04
1-3	1.24	1.11	1.24	1.11	1.16	1.04	1-3	1.04
1-4	1.24	1.11	1.24	1.11	1.16	1.04	1-4	1.04
1-5	1.24	1.11	1.24	1.11	1.16	1.04	1-5	1.04
2-3	1.07	1.11	1.07	1.11	1.16	1.04	2-3	1.04
2-4	1.07	1.11	1.07	1.11	1.16	1.04	2-4	1.04
2-5	1.07	1.11	1.07	1.11	1.16	1.04	2-5	1.04

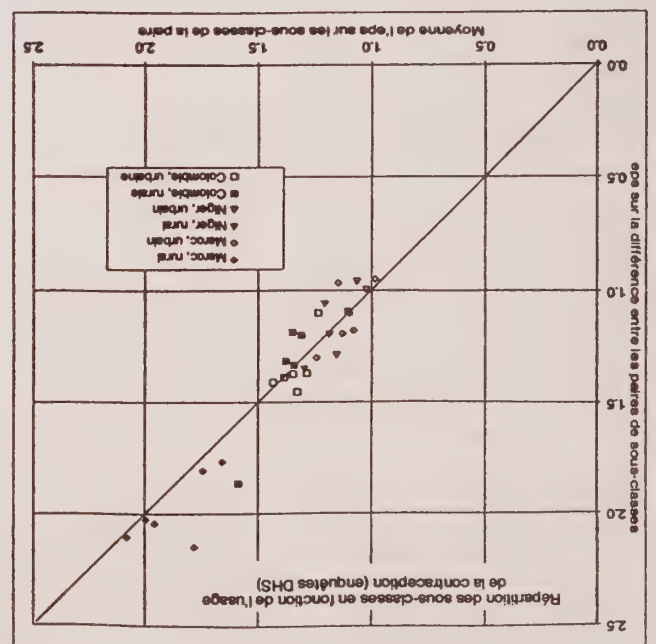


Figure 1. Comparaison de $\text{eps}(P_i - P_j)$ à la moyenne de $\text{eps}(P_i)$ et de $\text{eps}(P_j)$ pour diverses catégories d'usage de la contraception*. Illustration de six populations tirées des enquêtes sur la démographie et la santé des populations (DHS).

* 1 = aucune méthode de contraception
2 = usage des méthodes traditionnelles uniquement
3 = utilisation d'une méthode moderne "réversible"
4 = stérilisation.

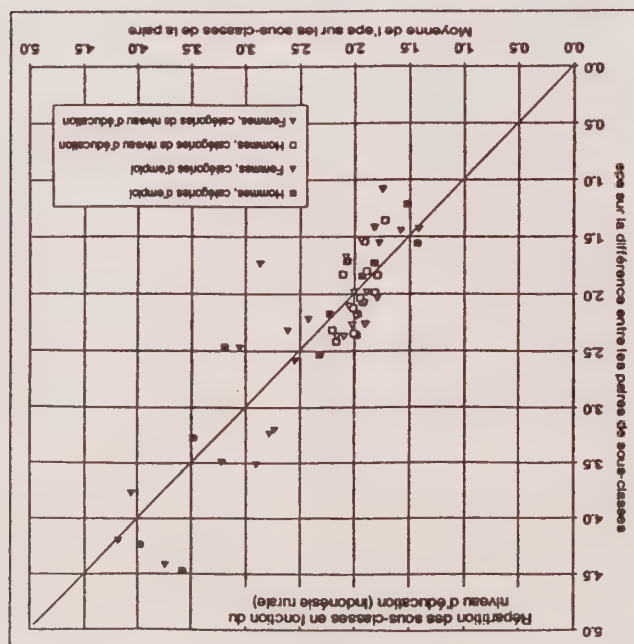


Figure 2. Comparaison de $\text{eps}(P_i - P_j)$ à la moyenne de $\text{eps}(P_i)$ et de $\text{eps}(P_j)$ pour les diverses catégories d'emploi et de niveau d'éducation par sexe. Illustration d'un recensement.

Catégories	$i - j$	P_i	P_j	Moyenne	$(P_i - P_j)$
Eps pour					

B. Position sur l'avortement

1-2	1.27	.97	1.12	1.07	1.00	.98
1-3	1.27	1.28	1.28	1.27	1.28	1.32
1-4	1.27	1.31	1.29	1.29	1.36	1.08
2-3	.97	1.28	1.12	1.12	1.08	1.16
2-4	.97	1.31	1.14	1.14	1.30	1.32
3-4	1.28	1.31	1.30	1.30	1.21	1.20
Moyenne	1.17	1.24	1.21	1.21		

D. Problèmes dans le pays

1-2	1.07	.94	1.00	1.07	1.00	.98
1-3	1.07	1.04	1.05	1.07	1.09	1.09
1-4	1.07	.93	1.00	1.07	1.12	1.01
2-3	.94	1.04	.99	.94	.85	.82
2-4	.94	.93	.93	.93	.85	.82
3-4	1.04	.93	.98	1.04	.98	.98
Moyenne	1.02	.97	.99	.97		

1-2	1.21	1.42	1.32	1.10	1.32	1.07
1-3	1.21	2.32	1.77	1.36	2.02	1.10
1-4	1.21	1.50	1.19	1.87	1.17	1.18
2-3	1.42	2.32	1.42	1.87	1.93	1.17
2-4	1.42	1.50	1.46	1.57	1.27	1.30
3-4	1.42	1.18	1.30	1.91	1.27	1.21
3-5	2.32	1.18	1.75	2.04	1.19	1.55
4-5	1.50	1.18	1.34	1.19	1.19	1.55
Moyenne	1.56	1.53	1.54	1.54		

C. Appui à Reagan

1-2	1.32	1.10	1.21	1.07	1.26	1.07
1-3	1.32	.86	1.09	1.26	1.09	1.07
1-4	1.32	1.40	.98	1.50	1.38	1.09
2-3	1.10	.86	1.10	.96	1.29	1.09
2-4	1.10	1.48	1.17	1.38	1.17	1.21
3-4	.86	1.48	1.17	1.38	1.17	1.21
Moyenne	1.17	1.21	1.19	1.21		

La National Election Study de 1986 réalisée par le Institute for Social Research de la University of Michigan ($n = 2,135$)

Tableau 1

4. RÉSULTATS EMPIRIQUES POUR $EPS(P_i - P_j)$

Figures:

1. Les études sur la démographie et la santé réalisées au Maroc, au Niger et en Colombie par MACRO International.
2. La portion du recensement indonésien portant sur la strate rurale de Java (données inédites).

Nous relevons en particulier dans ces études les résultats EMPIRIQUES suivants:

- 1) D'abord et avant tout: les effets du plan de sondage sur les différences $EPS(P_i - P_j)$ ne sont habituellement PAS MOINS IMPORTANTS que les $EPS(P_i)$ sur les proportions elles-mêmes, et $EPS(P_i - P_j) \approx 0.5 [EPS(P_i) + EPS(P_j)]$ dans tous les cas. Ils varient de concert avec l'importante variation des valeurs d'eps entre les variables, et avec les variations moins importantes observées entre les paires de catégories d'une même variable. Les chercheurs qui négligent les eps font l'erreur commune de sous-estimer les erreurs d'échantillonnage propres aux statistiques d'enquêtes par grappes. Outre l'intérêt qu'elle présente, cette observation fournit un modèle utile de déduction puisque les trois autres sources de variation – entre les variables, entre les catégories de chaque variable et erreurs d'échantillonnage des statistiques individuelles – sont toutes plus importantes.

- 2) On peut obtenir ces résultats avec les 14 ensembles de données d'enquêtes des tableaux et des graphiques, et en donner une illustration au tableau 1. Noter que les eps varient essentiellement entre 1.00 pour la variable D (problèmes dans le pays) et 2.32 pour la variable A (religion), ce qui nous donne une valeur d'eps² = $2.32^2 = 5.38$. Il est rassurant de constater que notre règle empirique (1) se vérifie pour toute la gamme des valeurs. On observe fréquemment une telle variation entre les variables d'un même échantillon; ceci devrait nous convaincre de cesser d'utiliser une moyenne commune pour tous les eps d'un échantillon (Verma et Le 1995; Kish 1995).

- 3) Des erreurs d'échantillonnage sont également présentes dans le calcul des eps. Il semble que seuls les statisticiens qui ont une bonne expérience du traitement des erreurs d'échantillonnage et des effets du plan de sondage puissent avoir une idée de l'importance de ces erreurs. Ces erreurs risquent d'être les principales responsables

À défaut d'un solide fondement théorique ou mathématique permettant d'accorder une préférence à l'une ou l'autre des conjectures énumérées plus haut, les résultats empiriques sur les valeurs d'eps ($P_i - P_j$) deviennent essentiels puisqu'ils permettent d'établir un rapport entre ces valeurs et les valeurs calculées d'eps (P_i). Ceci rappelle nos hypothèses plus familières concernant $EPS(P_i) = \sqrt{1 + \rho h} [b - 1]$, leur valeur dépend de plusieurs facteurs influant sur ρh , le coefficient de corrélation intracasse, ainsi que de la taille moyenne de la grappe b (Kish 1965, 5.4, 8.2). Les valeurs d'eps (P_i) varient beaucoup d'un sondage à l'autre, ainsi que d'une variable à l'autre dans un sondage particulier (Kish, Groves et Krotki 1976; Verma, Scott et O'Muircheartaigh 1980; Verma et Le 1995). Toutefois, les études empiriques portant sur les erreurs d'échantillonnage observées dans diverses enquêtes constituent, pour les statisticiens d'enquête, une source utile de connaissances qui leur permettent également d'établir des rapports entre les valeurs d'eps de statistiques complexes et celles d'eps (P_i) plus simples (Kish 1995; Rao et Wu 1985; Rao et Scott 1987). De même, pour en savoir plus sur le rapport entre $EPS(P_i - P_j)$ et $EPS(P_i)$, nous pouvons utiliser les résultats empiriques issus de plusieurs variables et de plusieurs enquêtes.

Dans un premier essai portant sur cette question, nous présentons les données tirées de quatorze enquêtes réalisées dans une vaste gamme de situations. Onze de ces enquêtes présentent cinq ensembles de données (figures 1 et 2 et tableaux 1 à 3) et portent sur les différences de catégories apparues tirées d'enquêtes uniques (type A). Trois ensembles de résultats (tableaux 1 à 3) proviennent d'enquêtes sociales, et les deux autres (figures 1 et 2) proviennent d'enquêtes sur la démographie et la santé des populations. Finalement, trois autres ensembles comportant chacun deux vagues de données sont fondés sur deux nouvelles séries d'entrevues menées auprès des mêmes répondants (tableaux 4, 5 et 6), et représentent les plans de comparaison du type B.

Tableaux:

1. La National Election Study de 1986, réalisée par le Institute for Social Research de la University of Michigan, $n = 2,135$.

2. La National Education Longitudinal Study (NELS) de 1988, réalisée par le National Opinion Research Center de la University of Chicago, $n = 24,355$.

3. La National Longitudinal Study of Labor Market Experience of Youth, réalisée par le National Opinion Research Center de la University of Chicago, $n = 5,857$.

4. Les National Election Studies Panels de 1990 et de 1992, réalisés par le Survey Research Center, Institute for Social Research, Ann Arbor, Michigan 48106.

5. Les Panel Study of Income Dynamics de 1983 et de 1987, réalisés par le Survey Research Center.

6. Les études Americans' Changing Lives de 1986 et de 1989 réalisées par le Survey Research Center.

3. ERREURS D'ÉCHANTILLONNAGE ET EFFETS DU PLAN DE SONDAGE

Pour les échantillons aléatoires simples de taille n , on peut facilement montrer (Kish 1965, 12.10) que

$$\text{var}(p_2 - p_0) = \left[\frac{(1-f)n}{(n-1)} \right] [p_2 + p_0 - (p_2 - p_0)^2] / n.$$

La plupart des exemples trouvés et utilisés sont tirés de grands échantillons d'enquêtes où la valeur de $(1-f)$ peut être ignorée. Il convient de noter que pour la variance des éléments

$$p_2 + p_0 - (p_2 - p_0)^2 = p_2q_2 + p_0q_0 + 2p_2p_0,$$

où le dernier terme $\text{cov}(p_2, p_0) = -p_2p_0$ désigne la covariance découlant de ce que p_2 et p_0 sont des parties concurrentes du même échantillon, plutôt que les proportions d'échantillons indépendants. Le carré de la différence de proportions $(p_2 - p_0)^2$ sera habituellement un terme de correction petit; en l'excluant, nous obtiendrons l'équivalent de la variance $(p_2 + p_0)/n$ de deux échantillons de Poisson indépendants. On se rappellera en outre (Kish 1965, 12.10) que:

Le test de chi carré a été utilisé pour certains de ces problèmes traités séparément (Cochran 1950; Mosteller 1952; McNemar 1962, p. 225). Il s'agit essentiellement de $(n_2 - n_0)^2 / (n_2 + n_0)$, c'est-à-dire du carré de la différence divisé par la variance, en vertu de l'hypothèse nulle $n_2 = n_0$. On applique exactement les théories disponibles pour les tests d'hypothèses nulles dans les petits échantillons, y compris la "correction de Yates", toutes fondées sur l'hypothèse d'un échantillonnage aléatoire simple. Toutefois, le traitement de ces problèmes dans de grands échantillons avec des moyennes estimatives et des écarts-types adéquats présente d'énormes avantages. Premièrement, au lieu de nous limiter à tester les hypothèses nulles, nous pouvons faire des déductions fondées sur les intervalles de probabilité $(p_2 - p_0) \pm t_{p, d.f.} (p_2 - p_0)$. Deuxièmement, les formules des écarts-types d'échantillons complexes peuvent s'appliquer directement à la moyenne $(p_2 - p_0)$. Troisièmement, la structure logique de cette statistique ($p_2 - p_0$) s'observe plus facilement dans son application à plusieurs problèmes distincts.

Les proportions corrélées proviennent habituellement de données tirées d'enquêtes complexes, et les calculs de la variance devraient être appropriés pour le plan de sondage. On peut adopter les formules de la variance propres aux échantillons complexes stratifiés, mais la formule directe comporte huit termes (Kish 1965, 12.10.3). Au lieu de cela, il est plus pratique de traduire le problème en une variable trichotomique avec des valeurs de $-1, 0, +1$,

comme dans le plan 2 de la section 2. Les calculs de la section 4 utilisent cette transformation.

Les comparaisons entre les variables et entre les échantillons peuvent alors être facilitées par le recours aux effets du plan de sondage:

$$\text{eps}^2(p_2 - p_0) = \frac{\text{variance calculée de } (p_2 - p_0)}{[p_2 + p_0 - (p_2 - p_0)^2] / n}.$$

Il convient de s'attarder quelque peu sur les limites de l'utilisation des eps en guise d'outils pour les approximations robustes. Ils sont utiles pour les échantillons en grappes et les échantillons à plusieurs degrés qui utilisent les sélections primaires pour le calcul des erreurs d'échantillonnage. Toutefois, nous avons voulu éviter le problème des échantillons pondérés parce que leur traitement serait trop spécifique et peut-être trop complexe. La pondération pour l'absence de réponse ne serait pas importante pour le rapport de $\text{eps}(p_i - p_j)$ sur $\text{eps}(p_i)$. Toutefois, la pondération des inégalités majeures des probabilités de sélection appelle des traitements particuliers. Néanmoins, la déduction et l'expérience portent à conclure que les valeurs d'eps sont moins sensibles aux poids que les variances et les moyennes elles-mêmes. Par ailleurs, nous présumons que les rapports que nous avons observés entre les valeurs d'eps ($p_i - p_j$) et d'eps (p_i) se maintiendront avec les données pondérées, à condition que ces dernières ne soient pas extrêmes ni pathologiques.

Il serait utile de déterminer un rapport approximatif, mais fiable, d'eps ($p_i - p_j$) sur $\text{eps}(p_i)$ et $\text{eps}(p_j)$ afin de pouvoir déduire la valeur du premier, peu aisée à calculer, à partir des valeurs des seconds que l'on peut obtenir plus facilement. Plusieurs conjectures de échange peuvent être dérivées mathématiquement ni exclues.

1. $\text{Eps}(p_i - p_j) = 1$, l'absence d'effet du plan de sondage, a été implicitement présuée dans les cinq publications citées à la section 1.

2. $\text{Eps}(p_i) > \text{eps}(p_i - p_j) > 1$, désignant des effets persistants, mais moins importants que les $\text{eps}(p_i)$ pour les proportions. C'est ce qui se produit dans le cas des "classes croisées" et de leurs comparaisons (Kish 1987, 7.1). Cette hypothèse a également paru raisonnable à plusieurs statisticiens d'expérience que nous avons consultés.

3. $\text{Eps}(p_i - p_j) = [\text{eps}(p_i) + \text{eps}(p_j)] / 2$ nous a semblé être une bonne approximation pour l'ensemble de nos données pour diverses populations et divers plans de sondage. Cette hypothèse nous paraît raisonnable puisqu'elle est issue de la différence entre des valeurs individuelles de p_i peuvent s'appliquer de façon similaire à la variable issue de la différence $(p_i - p_j)$ entre deux de ces valeurs.

4. Des résultats incohérents seraient également possibles, mais peu utiles puisqu'ils empêcheraient toute déduction.

citer à titre d'exemple la différence entre les proportions de gens qui "interdiraient tous les essais nucléaires" et ceux qui "souhaitent un désarmement nucléaire complet", ou entre ceux qui "obligeraient les Irakiens à quitter le Koweït" et ceux qui voudraient "remplacer Saddam" (Wild et Seber 1993). Toutefois, les deux catégories pourraient également être tirées de deux *sondages différents* portant sur les mêmes n cas, si on avait affaire par exemple à une vérification de la qualité, à des observations provenant de deux bases de sondage ou aux deux vagues d'un même échantillon. Ces situations ressemblent à celles examinées aux points (4) et (5).

Comparaisons du type B

4. Les expressions test-retest et avant-après qualifient des plans dans lesquels les mêmes sujets font l'objet de deux observations. Les réponses dichotomiques $n_2 = n_{10}$ désignent alors le nombre de changements positifs et $n_{01} = n_{01}$ le nombre de changements positifs et $n_{11} + n_{00}$ la somme des positifs et des négatifs inchangés. On dénote respectivement les réponses positives et négatives par 1 et 0, et les première et deuxième vagues par l'ordre de l'indice. La différence $(n_{10} + n_{11}) - (n_{01} + n_{11}) = n_{10} - n_{01} = n_2 - n_0$ représente une mesure du changement entre les positifs pour les deux observations, et $p_2 - p_0 = n_2/n - n_0/n$ représente la mesure du changement des proportions (McNemar 1949; Cochran 1950; Mosteller 1952).

5. Les paires appariées de n paires de sujets peuvent également être traitées comme une généralisation du plan test-retest (Mosteller 1952). Par exemple, n paires de sujets randomisés peuvent représenter le traitement expérimental par rapport au traitement témoin; ou n paires de garçons ou de filles par rapport aux variables de contrôle. Le traitement statistique $(p_{10} - p_{01})$ des n paires de sujets appariés est le même que celui des n paires de traitements du même groupe de n sujets (4).

La similitude des traitements statistiques de ces cinq plans de deux types distincts est pratique; nous présentons les résultats empiriques obtenus pour les deux types. "Elle a également une valeur heuristique qu'on a ignorée dans des publications récentes (Scott et Seber 1983; Wild et Seber 1993). Les résultats des tests de chi carré pour les types 4 et 5 ont été publiés depuis longtemps (McNemar 1949; Cochran 1950; Mosteller 1952) et la similitude des cas des catégories 1, 2 et 3 a également été démontrée" (Kish 1965, 12.10). (Kish faisait erreur lorsqu'il a parlé de "trinomies et de binômes appariés" pour désigner les "trichotomies et les dichotomies appariées" puisque ces termes s'appliquent uniquement aux échantillons IID.) Tous ces traitements portent sur les différences des proportions p_i fondées sur les variables de dénombrement n_i . Les extensions aux différences corrélées $(y_i - y_j)$ pour d'autres variables sont envisageables, mais elles sortent du cadre de la présente étude. On pourrait citer à titre d'exemple pratique la différence entre les parts en dollars (pas seulement les nombres n_i) entre deux marques d'automobiles sur un total de $\sum y_i$ ventes.

Dans la section 2, nous décrivons les cinq problèmes appariés (plans de sondage) au sujet desquels nous décrivons les erreurs d'échantillonnage à la section 3. Dans la section 4, nous examinons la constatation empirique exposée dans les tableaux. Dans la section 5, nous plaçons nos résultats dans le contexte des travaux antérieurs portant sur les valeurs d'eps pour les sous-classes et leurs différences.

2. STATISTIQUES SEMBLABLES POUR LES CINQ PLANS DE SONDAGE

On a montré que cinq plans de sondage (problèmes) appartenant à deux types distincts peuvent être traités à l'aide des mêmes statistiques simples (Kish 1965, Section 12.10). Pour notre présentation à la fois empirique et simple, nous utilisons des symboles qui dénotent les valeurs d'échantillons (eps, p_i et n_i), même si à l'occasion, l'usage de majuscules pour les valeurs de la population totale serait plus approprié.

La différence de proportions $p_2 - p_0 = n_2/n - n_0/n$ dénote l'estimation souhaitée, où $n = n_0 + n_1 + n_2 + \dots$ n_k désigne la taille de l'échantillon, avec n unités choisies et pondérées également. Par ailleurs, en vertu des hypothèses de l'échantillonnage aléatoire simple, la variance de $(p_2 - p_0)$ est $(1 - f) [p_2 + p_0 - (p_2 - p_0)^2] / (n - 1)$.

Comparaisons de type A

1. La différence entre deux catégories $(n_2 - n_0) / n = (p_2 - p_0)$ d'une polytomie peut représenter la préférence manifestée pour deux partis parmi plusieurs (k) dans un sondage sur les intentions de vote, pour deux marques d'automobiles dans une étude de marchés, pour deux types d'attitudes, d'opinions ou de comportements concernant une variable quelconque, etc. Les autres choix $(k - 2)$ sont pris en compte dans p_1 et sont ignorés dans l'évaluation de la différence. (Voir à ce propos Scott et Seber 1983)

2. On peut attribuer les valeurs de rang de $-1, 0, +1$ (ou $0, 1, 2$ ou $c, c + 1, c + 2$) à une variable trichotomique ordonnée sans métrique, et les assimiler à une forme simple de la différence entre deux catégories. Cette forme est particulièrement utile pour le calcul des erreurs d'échantillonnage puisque les cinq plans de sondage peuvent par exemple utiliser $-1, 0, +1$ à titre de variable de calcul transformée.

3. La différence de proportions de deux variables différentes $(x$ et $y)$ peut être traitée comme en (1) et en (2). Ne définir comme positifs en x (succès) que les éléments qui sont positifs en x , mais non en y , de manière que $n_{10} = n(x_1, y_0)$. Définir de la même manière comme positifs en y les $n_{01} = n(x_0, y_1)$. Ainsi, $(n_{10} - n_{01}) / n = (p_x - p_y)$ représentera la différence nette dans la proportion des éléments positifs en x et y . Les éléments qui sont positifs ou négatifs à la fois en x et y ne comptent pas dans les différences. Nous obtenons ainsi un cas de trois catégories comme en (1) et en (2). On peut

Effets du plan de sondage sur les $(P_i - P_j)$ corrélés

LESLIE KISH, MARTIN R. FRANKEL, VIJAY VERMA et NIKO KACIROTI¹

RÉSUMÉ

En nous fondant sur 14 enquêtes menées dans six pays, nous présentons la constatation empirique de l'existence et de l'ampleur des effets du plan de sondage (eps) pour cinq plans appartenant à deux types principaux. Le premier type a trait à $\text{eps}(P_i - P_j)$, la différence de deux proportions d'une variable polynomique de trois catégories ou plus. Le deuxième type utilise les tests de chi carré pour l'analyse des différences entre deux échantillons. Nous montrons que pour toutes les variables et pour tous les plans, $\text{eps}(P_i - P_j) \equiv [\text{eps}(P_i) + \text{eps}(P_j)]/2$ constituent de bonnes approximations. Ces résultats sont *empiriques*, et les exceptions proviennent qu'il ne peut s'agir de simples inégalités analytiques. Il convient de signaler que ces résultats restent valables malgré les grandes variations des valeurs d'eps entre les variables et entre les catégories d'une même variable. Ils montrent en outre la nécessité d'utiliser des méthodes de traitement adaptées aux échantillons d'enquêtes pour l'analyse des données d'enquête, même lorsqu'on a affaire à des statistiques analytiques. En outre, ils permettent d'utiliser des approximations d'eps ($P_i - P_j$) tirées des valeurs plus facilement accessibles d'eps (P_i).

MOTS CLÉS: Effets du plan de sondage; échantillonnage d'enquête; erreurs d'échantillonnage.

1. EFFETS DU PLAN DE SONDAGE SUR LES STATISTIQUES ANALYTIQUES

Nous nous penchons sur l'existence et sur l'ampleur des effets du plan de sondage sur certaines statistiques analytiques spéciales, en nous servant de données tirées d'échantillons d'enquêtes. Notre étude s'inspire à la fois de la méthode et de l'empirisme; elle s'appuie sur des données venant de plusieurs enquêtes aux variables différentes et provenant de populations contrastantes, et s'expose donc aux risques que présentent les résultats empiriques inhérents. On dit et on écrit souvent que l'échantillonnage probabiliste, bien que nécessaire à la réalisation d'enquêtes descriptives, n'est pas nécessaire pour les enquêtes analytiques. Dans une section intitulée "Four Obstacles to Representation in Analytic Studies", l'un de nous écrit qu'en plus des quatre obstacles concrets à la représentation auxquels il est fait allusion, il en existe un autre plus artificiel: le refus d'admettre la nécessité de la représentation (Kish 1987, section 2.7). Les études sur l'échantillonnage montrent que les méthodes de sélection probabiliste complexes, en particulier l'échantillonnage par grappes, n'ont pas d'effet appréciable sur les statistiques descriptives (p. ex., les moyennes et les coefficients de régression), mais qu'elles peuvent influencer d'une façon extrêmement marquée sur les statistiques déductives comme les intervalles de confiance et les tests de signification (Kish et Frankel 1974). Les effets du plan de sondage (eps) sont définis par la formule $\text{eps}^2 = \text{variance réelle/variance aléatoire simple}$, pour une même taille d'échantillon n (les deux variances étant estimatives). On a observé des valeurs d'eps supérieures à 1 pour des erreurs d'échantillonnage concernant non seulement les moyennes, mais également des statistiques analytiques comme les écarts de moyennes (et les tests de chi carré), les coefficients de régression, etc. Il est vrai

qu'on a observé des réductions et des différences considérables des valeurs d'eps pour certaines statistiques analytiques. Les variations des valeurs d'eps ne sont pas une simple conséquence mathématique inévitable du choix du plan de sondage que l'on peut déduire une fois pour toutes. Elles ont un contenu empirique et doivent donc être reproductibles dans le cadre d'études empiriques (Kish et Frankel 1974; Kish 1987, 7.1; Kish 1965, 14.1-14.2; Rao et Wu 1985; Scott et Holt 1982; Skinner, Holt et Smith 1989). Dans le présent article, nous examinons l'incidence possible et l'importance des effets du plan de sondage sur un ensemble de statistiques apparentées qui n'ont fait l'objet d'aucune étude antérieure. Il se trouve que, dans de nombreux articles, des auteurs reconnus à bon droit ont simplement présumé que le choix du plan de sondage était sans effet, sans fournir aux lecteurs les mises en garde appropriées, et que cela n'a soulevé aucune objection de la part des examinateurs. Nous vérifierons donc si les eps sont réduits ou éliminés pour cet ensemble de statistiques analytiques (Cochran 1950; Mosteller 1952; Scott et Seber 1983; Seber et Wild 1993). Par ailleurs, nous proposerons explicitement ce à quoi nous avons déjà fait allusion auparavant, à savoir que l'existence d'effets de sondage importants justifie largement le recours à la sélection probabiliste. Il serait difficile d'imaginer un modèle de distribution de population où le plan de sondage serait sans importance (ou sans intérêt) tout en produisant des effets considérables. Toutefois, à l'inverse, l'absence d'eps est nécessaire, quoique non suffisante, pour justifier la décision de ne pas recourir à la sélection probabiliste. Cette proposition donne du poids à notre étude qui met en rapport les valeurs d'eps ($P_i - P_j$) des statistiques analytiques d'une part, et celles d'eps (P_i et d'eps (P_j) pour deux des nombreuses catégories de la même variable d'autre part.

¹ Leslie Kish, ISR, University of Michigan, Ann Arbor MI 48106, U.S.A.; Martin R. Frankel, NORC et City University of New York; Vijay Verma, University of Essex, Colchester, CO4 3SQ, U.K.; Niko Kaciroti, Institut de statistiques, Tirana, Albanie.

- HICKS, S., et FETTER, M. (1993). An evaluation of robust estimation techniques for improving estimates of total hogs. *Proceedings of the International Conference on Establishment Surveys*. Alexandria, Virginia: American Statistical Association, 385-389.
- HIDIROGLOU, M.A. (1987). The construction of a self-representing stratum of large units in survey design. *The American Statistician*, 40, 27-31.
- HUBER, P.J. (1981). *Robust Statistics*. New York: John Wiley.
- LEE, H. (1994). Outliers in Survey Data. Document de Statistique Canada.
- RIVEST, L.-P. (1994). Some sampling properties of winsorized means for skewed distributions. *Biometrika*, 81, 373-384.
- RIVEST, L.-P. (1993a). Winsorization of survey data. *Bulletin de l'Institut International de Statistique, Actes de la 49^{ème} Session*, livre 2, 73-89.
- RIVEST, L.-P. (1993b). Winsorization of survey data. *Proceedings of the International Conference on Establishment Surveys*. Alexandria, Virginia: American Statistical Association, 396-401.
- SEARLS, D.T. (1966). An estimator which reduces large true observations. *Journal of the American Statistical Association*, 61, 1200-1204.
- THISTED, R.A. (1988). *Elements of Statistical Computing*. New York: Chapman and Hall.
- VON MISES, R. (1964). *Selected Papers of Richard von Mises*. Volume II. Providence: American Mathematical Society.

$$\psi(R) - E(Y) \frac{n-1}{n} > \int_{\psi(R)}^{\infty} [1 - F_Y(x)] dx. \quad (A.1)$$

Selon l'inégalité de Jensen, $E(Y) = E[\psi(X)] > \psi[E(X)]$. Ainsi, en utilisant (2.3), le membre de gauche de l'équation (A.1) est inférieur ou égal à

$$\frac{R - E(X) \psi(R) - \psi[E(X)]}{R - E(X)} \frac{n-1}{n} > \frac{n-1}{R - E(X)} \psi'(R) = \int_R^{\infty} [1 - F_X(y)] dy \cdot \psi'(R)$$

où ψ' est la dérivée de ψ . Comme ψ' augmente, le membre de gauche de l'inégalité ci-dessus est inférieur ou égal à:

$$\int_{-\infty}^R \psi'(y) [1 - F_X(y)] dy = \int_{-\infty}^{\psi(R)} [1 - F_Y(x)] dx.$$

ce qui prouve la validité de (A.1).

Preuve de la proposition 2 Le résultat suivant obtenu en appliquant les théorèmes 2.7.5 et 2.7.11 de Galambos (1987) à la distribution $F(z^{p+1})$ est largement mis à contribution. Si le maximum de l'échantillon de la distribution $F(x)$ tend vers $H_{3,0}(x)$, tous les moments de F existent et

$$\int_{-\infty}^x y^p [1 - F(y)] dy \sim \frac{g(x)}{[1 - F(x)] x^p} \quad (A.2)$$

où $g(x) \sim h(x)$ signifie que $g(x)/h(x)$ tend vers 1 à mesure que x tend vers l'infini. À l'aide de la formule (A.2), on obtient $R(F, n)$ en résolvant

$$\frac{R - \mu}{R - 1} = \frac{g(R)}{[1 - F(R)] (1 + o(1))}.$$

Soit $R = F^{-1}(1 - a/n)$, alors, jusqu'à la valeur $(1 + o(1))$, l'équation ci-dessus devient

$$a = g \left[F^{-1} \left(1 - \frac{n}{a} \right) \right] F^{-1} \left(1 - \frac{n}{a} \right). \quad (A.3)$$

Soit $a_0 = g[F^{-1}(1 - 1/n)] F^{-1}(1 - 1/n)$ et $a_1 = g[F^{-1}(1 - a_0/n)] F^{-1}(1 - a_0/n)$. Puisque pour les grandes valeurs de x , $g(x)$ augmente, $a_0 > a_1$ et la solution de (A.3) tombe dans l'intervalle (a_1, a_0) . Pour prouver le résultat, il suffit de montrer que a_1/a_0 tend vers 1 à mesure que n tend vers l'infini.

Puisque $g(x) = f(x)/[1 - F(x)]$, on peut poser

BIBLIOGRAPHIE

- BARLOW, R.E., et PROSCHAN, F. (1981), *Statistical Theory of Reliability and Life Testing*. Silver Spring MD: To Begm With.
- CHAMBERS, R.L., et KOKIC, P.N. (1993). Outlier robust sample survey inference. *Bulletin de l'Institut International de Statistique*, Actes de la 49^{ème} Session, livre 2, 54-72.
- ERNST, L.R. (1980). Comparison of estimators of the mean which adjust for large observations. *Sankhyā C*, 42, 1-16.
- FELLER, W. (1971). *An Introduction to Probability Theory and its Applications*. Volume II. Deuxième édition. New York: Wiley.
- FULLER, W.A. (1991). Simple estimators for the mean of skewed populations. *Statistica Sinica*, 1, 137-158.
- FULLER, W.A. (1993). Estimators for long-tailed distributions. *Bulletin de l'Institut International de Statistique*, Actes de la 49^{ème} Session, livre 2, 39-54.
- GALAMBOS, J. (1987). *The Asymptotic Theory of Extreme Order Statistics*. Deuxième édition. Malabar FL: Krieger.
- GNEDENKO, B.V. (1962). *The Theory of Probability*. New York: Chelsea.
- GLASSER, G.J. (1962). On the complete coverage of large units in a statistical study. *Revue de l'Institut International de Statistique*, 30, 28-32.

Preuve de la proposition 3 Si le maximum de l'échantillon de la distribution $F(x)$ tend vers $H_{1,\alpha}(x)$, F satisfait alors aux propriétés suivantes (Feller 1971, p. 281):

pour toute valeur de p telle que $\alpha - p - 1 \geq 0$. Avec la formule (A.4), $R(F, n)$ s'obtient en résolvant $F(R) = 1 - [\alpha - 1 + o(1)]/n$. Ceci conduit à l'approximation de $R(F, n)$. Pour calculer l'approximation de $\text{MSE}(\bar{X}_R)$, on utilise (A.4) avec $p = 1$.

où la seconde expression est obtenue en intégrant par parties. Puisque $ig'(t)/g(t)$ est inférieur à c , on obtient $a_0 > \exp(a_0 - a_1)a_0^{-c}$. Si a_1/a_0 ne tend pas vers 1, c 'est-à-dire par exemple que $a_1/a_0 < 1 - \epsilon < 1$ pour une séquence infinie de tailles d'échantillons, l'inégalité précédente signifie que $a_0^{1+c} > \exp(a_0\epsilon)$. Ce résultat est contradictoire puisque a_0 tend vers l'infini à mesure que n devient plus grand. L'approximation de $\text{MSE}(\bar{X}_R)$ s'obtient en utilisant la formule (A.2) avec $p = 2$.

$$a_0 = \exp \left[\int_{F^{-1}(1-1/n)}^{F^{-1}(1-a_0/n)} g(t) dt \right] = \exp \left[a_0 - a_1 - \int_{F^{-1}(1-1/n)}^{F^{-1}(1-a_0/n)} ig'(t) dt \right],$$

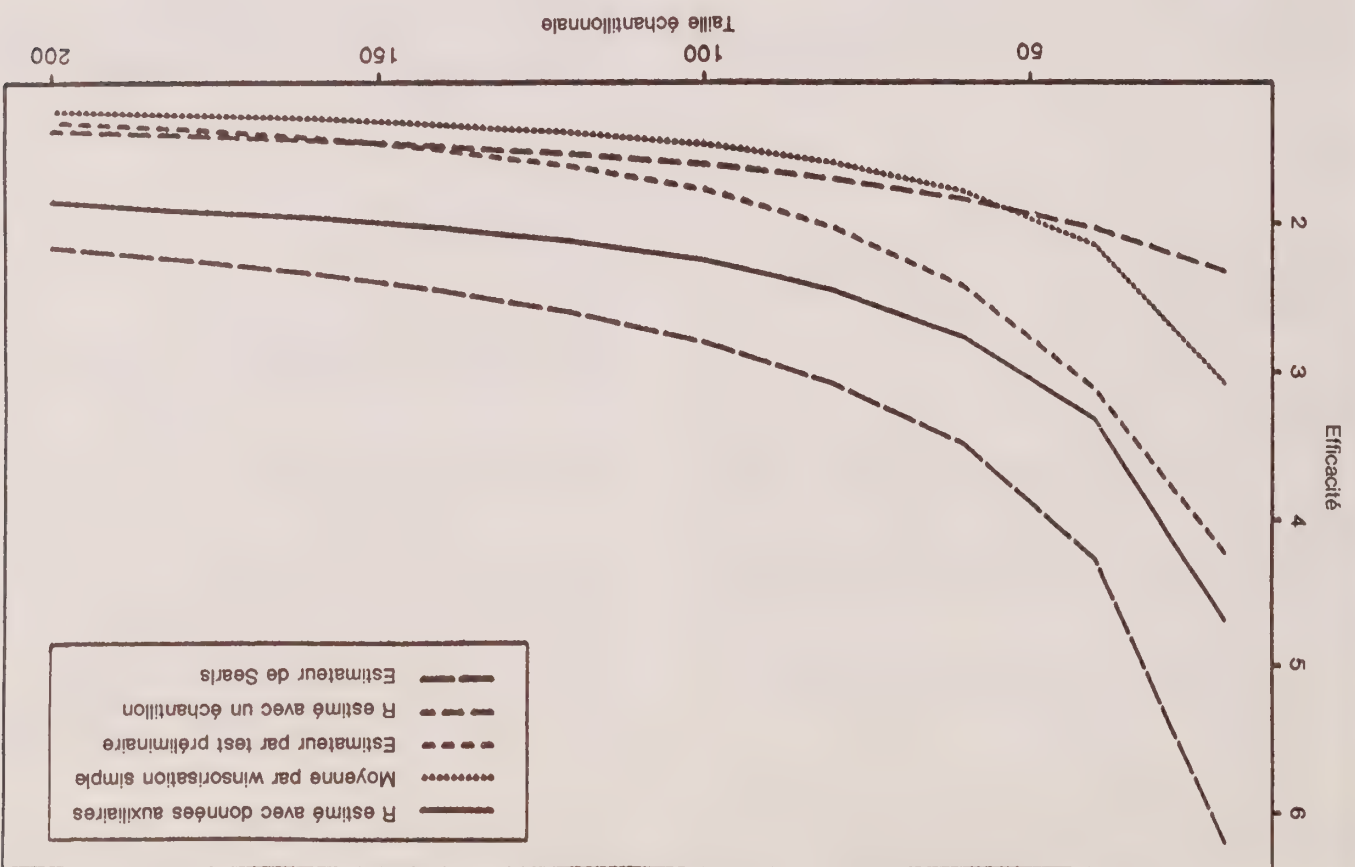


Figure 4. Efficacité de cinq estimateurs pour la moyenne du nombre de poulets.

avec les petits échantillons. Ainsi, comme nous l'avons montré dans la section 4, l'estimateur obtenu est très sensible aux valeurs aberrantes qui apparaissent parfois dans les petits échantillons. Cet estimateur n'est pas recommandé.

6. CONCLUSIONS

Un grand nombre de stratégies peuvent nous permettre de traiter les grandes valeurs qui apparaissent parfois lors des enquêtes. Si on dispose d'informations auxiliaires (par exemple, des données de recensement), on peut utiliser l'estimateur de Searls avec l'échantillonnage aléatoire simple, l'échantillonnage stratifié ou l'échantillonnage avec ppt. Comme les valeurs limites sont des constantes fixes, les estimateurs de l'erreur quadratique moyenne peuvent être dérivés à l'aide des formules (2.3) et (3.1). En l'absence d'informations auxiliaires, on peut utiliser la moyenne de la winsorisation simple et l'estimateur de test préliminaire de Fuller. Des travaux sont en cours afin de généraliser ces estimateurs avec les plans d'échantillonnage stratifié. Rivest (1994) propose un estimateur de l'erreur quadratique moyenne de la moyenne de la winsorisation simple:

$$v(\bar{X}_1) = \frac{1}{n} S^2 - \frac{1}{n^2} (X_n + X_{n-1} - 2\bar{X}_1)$$

$$(X_n - 3X_{n-1} + 2X_{n-2})$$

où S^2 désigne la variance de l'échantillon X et $X_n > X_{n-1} > X_{n-2}$ désignent les trois plus grandes valeurs de cet échantillon. Cet estimateur présente un léger biais dans les populations infinies. Toutefois, la couverture de l'intervalle de confiance normal est souvent bien en dessous du niveau nominal de $100(1 - \alpha)\%$, en particulier lorsque la distribution sous-jacente est asymétrique. De plus, amples recherches sont nécessaires pour obtenir des intervalles de confiance fiables pour les estimateurs de la moyenne des populations asymétriques.

REMERCIEMENTS

La présente recherche a bénéficié de l'aide financière du Conseil de recherches en sciences naturelles et en génie et du Fonds pour la formation des chercheurs et l'aide à la recherche du Québec.

ANNEXE 1

Preuve de la proposition 1 La supposition selon laquelle Y est plus asymétrique que X signifie qu'il existe une fonction convexe ψ telle que $\psi(X)$ et Y suivent la même distribution. Désignons par R la valeur $R(F_X, n)$. Pour prouver le résultat, il suffit de montrer que $\psi(R) < R(F_Y, n)$. Ceci équivaut à dire que

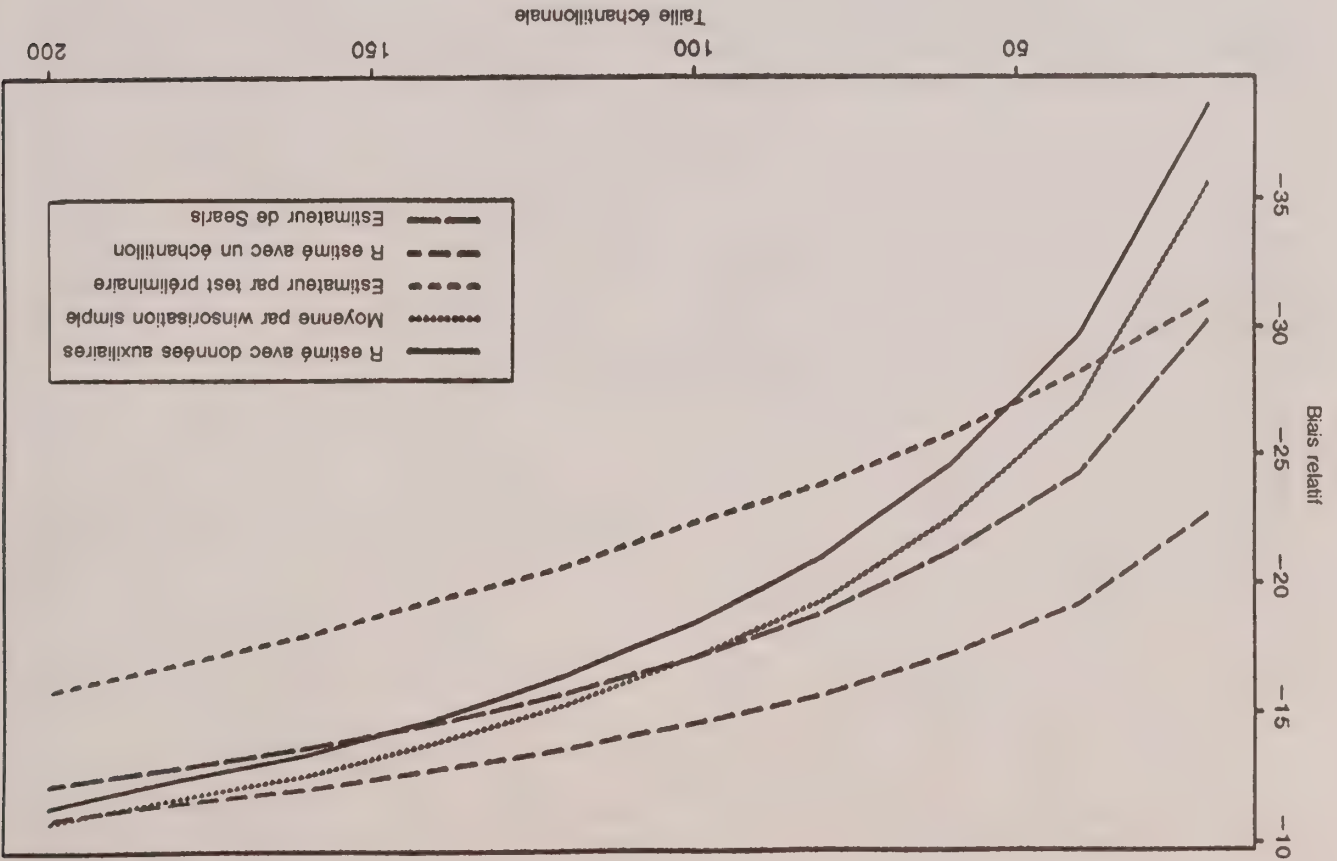


Figure 3. Biases relatifs de cinq estimateurs pour la moyenne du nombre de poulets.

Les cinq estimateurs considérés sont:

- L'estimateur winsorisé de Searls, \bar{X}_R , calculé comme si la distribution sous-jacente était connue;

- Un estimateur winsorisé où la valeur limite correspond à la deuxième valeur en importance d'un échantillon auxiliaire de taille $2n$; il s'agit d'un cas où on dispose d'informations auxiliaires limitées sur la distribution sous-jacente F (dans les simulations de Monte Carlo, chaque échantillon simulé possède son propre échantillon auxiliaire);

- La moyenne \bar{X}_1 de la winsorisation simple, présentée dans la section 4;

- Un estimateur winsorisé où R est estimé à partir de l'échantillon par la résolution de l'équation (4.3);

- L'estimateur d'un test préliminaire plus complet avec $j = 3$ (c.-à-d., le numérateur du test préliminaire utilise les trois plus grandes observations), T (le nombre total de points de données utilisés dans le test préliminaire) étant égal à $[4n^{1/2} - 10]$ et K_3 , la valeur limite, étant égale à 3.5. On fournit une description détaillée de cet estimateur dans Fuller (1991) et dans Rivest (1993a et b). Cet estimateur ne réduit les valeurs les plus grandes que lorsqu'un test statistique de détection des valeurs extrêmes donne des résultats significatifs.

Des trois estimateurs de la figure 4 qui ne dépendent pas temps dans la distribution de la variable à l'étude. compte des changements risquant d'être survenus avec le provenant des enquêtes antérieures normalisées pour tenir tillons auxiliaires pourraient être constitués de données (1993a). Dans le contexte d'un échantillonnage, les échan- d'une variance finie (pour en savoir plus, voir Rivest modélisée par la distribution d'une superpopulation assortie cace. Ceci reste vrai tant que la variable à l'étude peut être d'une information auxiliaire limitée est extrêmement effi- En outre, l'estimation de la valeur limite optimale à l'aide plus efficace que la moyenne de la winsorisation simple. prévoir le tableau 2, l'estimateur de Searls est beaucoup importantes de la figure 4. Premièrement, comme le laissait tillons. On peut par ailleurs tirer plusieurs conclusions risés sont importants, même dans le cas des grands échan- La figure 3 montre que les biais des estimateurs winso- des simulations de Monte Carlo, avec 100,000 répétitions. tats présentés dans les figures 1 et 2 ont été obtenus à l'aide calculés exactement. Pour les autres estimateurs, les résul- Les biais et les valeurs de l'efficacité de \bar{X}_R ont été

Tableau 2

Approximations du biais et de l'erreur quadratique moyenne de la moyenne de la winsorisation simple X_1 et de la moyenne winsorisée optimale de Searls, X_R , pour les distributions de Weibull et de Pareto ($\Gamma(\cdot)$ représente la fonction gamma)

WEIBULL		PARETO	
X_R	MSE	$\frac{\sigma^2}{n} - \frac{n^2}{(\log n)^{2\alpha}}$	$\frac{\sigma^2}{n} - \frac{(\gamma - 2)(\gamma - 1)^{2/\gamma} n^{2-2/\gamma}}{\gamma}$
	bias	$\frac{1}{(\log n)^\alpha} - \frac{n}{1}$	$\frac{1}{1} - \frac{(\gamma - 1)^{1/\gamma} n^{1-1/\gamma}}{1}$
X_1	MSE	$\frac{\sigma^2}{n} - \frac{2\alpha(\alpha - 1)(\log n)^{2\alpha-2}}{n^2}$	$\frac{\sigma^2}{n} - \frac{2\Gamma(1 - 2/\gamma)(\gamma - 1)^{2-2/\gamma}}{2\Gamma(1 - 2/\gamma)}$
	bias	$\frac{n}{\alpha(\log n)^{\alpha-1}} - \frac{\gamma n^{1-1/\gamma}}{\Gamma(1 - 1/\gamma)}$	

$$\frac{R - \bar{X}}{n - 1} = \frac{1}{n} \sum_{i=1}^n \max(X_i - R, 0) \quad (4.3)$$

pour fonction d'estimation de R . Cette méthode est contestable lorsque la distribution sous-jacente est fortement asymétrique, c'est-à-dire lorsque F satisfait l'hypothèse de la proposition 3. En moyenne, il n'y aura que des termes $\alpha - 1$ non nuls dans le membre de droite de l'équation (3). Ainsi, R sera déterminé, en moyenne, par les $\alpha - 1$ plus grandes valeurs et le maximum de l'échantillon aura la plus grande incidence sur R . Ceci rendra la valeur R extrêmement instable et, compte tenu des constatations faites concernant la figure 1, la valeur venant au deuxième rang pour sa grandeur deviendra probablement un estimateur plus adéquat de $R(F, n)$ que la solution de (3.3). Les simulations de Monte Carlo présentées dans la section 5 sont instructives à cet égard.

Le tableau 2 compare les approximations du biais et de l'erreur quadratique moyenne de la moyenne winsorisée de Searls, X_R , à celles de la moyenne de la winsorisation simple X_1 , obtenue en utilisant une valeur limite R égale à la deuxième plus grande observation. Rivest (1994) montre que ce choix de la valeur limite donne la moyenne winsorisée non paramétrique optimale. Il dérive également les approximations, pour grands échantillons, du biais et de l'erreur quadratique moyenne de X_1 qui sont présentées dans le tableau 2. Les expressions correspondantes pour X_R sont tirées des propositions 2 et 3. Dans le tableau 2, l'erreur quadratique moyenne de X_R est beaucoup plus petite que celle de X_1 . En fait, pour la distribution de Weibull, l'efficacité de X_R par rapport à X_1 pour les grands échantillons est égale à celle de X_R par rapport à X_1 . Ainsi, la winsorisation non paramétrique réduit l'erreur

5. COMPARAISONS DE MONTE CARLO DES ESTIMATEURS DE LA MOYENNE D'UNE DISTRIBUTION ASYMÉTRIQUE

Nous présentons ci-après les comparaisons de Monte Carlo de l'erreur quadratique moyenne et du biais de cinq estimateurs de la moyenne du nombre de poullets par segment d'une population comportant 2,000 unités (Fuller 1991). Le coefficient de variation est de 4.46. De plus amples comparaisons numériques des cinq estimateurs examinés dans la présente section pour d'autres distributions, finies ou infinies, sont présentées par Rivest (1993a et b).

Nous présentons ci-après les comparaisons de Monte Carlo de l'erreur quadratique moyenne et du biais de cinq estimateurs de la moyenne du nombre de poullets par segment d'une population comportant 2,000 unités (Fuller 1991). Le coefficient de variation est de 4.46. De plus amples comparaisons numériques des cinq estimateurs examinés dans la présente section pour d'autres distributions, finies ou infinies, sont présentées par Rivest (1993a et b).

4. APPROXIMATIONS DE L'EFFICACITÉ DE LA MOYENNE WINSORISÉE POUR LES GRANDS ÉCHANTILLONS

Pour la plupart des distributions, l'équation (2.3) définissant la valeur limite optimale n'a pas de solution explicite. Dans la présente section, nous obtenons des approximations directes de cette solution en utilisant la théorie des statistiques d'ordre extrême. Cette méthode nous permettra de calculer des approximations explicites de l'efficacité de la moyenne winsorisée optimale. Nous comparerons ensuite la méthode de winsorisation optimale de Searls à une méthode simple de winsorisation non paramétrique où la variable statistique la plus grande est remplacée par celle qui vient au second rang (Rivest 1994).

La forme de l'approximation de $R(F, n)$ dépend de la distribution limitante de la variable statistique d'ordre supérieur adéquatement normalisée, lorsque la taille de l'échantillon n tend vers l'infini. Pour les distributions tracées sur un axe positif, il n'existe que deux distributions limitatives possibles qui sont données par Galambos (1987, p. 53-54):

$$H_{1,\alpha}(x) = \exp(-x^{-\alpha}) \quad \text{pour } x > 0 \quad \text{et } \alpha > 0$$

et

$$H_{3,0}(x) = \exp[-\exp(-x)] \quad \text{pour } x \text{ dans } R.$$

Pour beaucoup des distributions utilisées dans l'analyse statistique des variables aléatoires positives (par exemple, les familles de Weibull et log-normale), le maximum de l'échantillon, adéquatement normalisé, tend vers $H_{3,0}(x)$. Les distributions dont le maximum de l'échantillon tend vers $H_{1,\alpha}(x)$ pour un certain $\alpha > 0$ sont caractérisées par des queues importantes. Pour de telles distributions, $1 - F(x)$ tend vers 0 à un taux de $O(x^{-\alpha})$. Les distributions de Pareto et les distributions de F appartiennent à cette classe.

Les distributions dont le maximum de l'échantillon tend vers $H_{3,0}(x)$ sont examinées les premières. La caractérisation suivante est tirée de von Mises (1964): le maximum de l'échantillon d'une distribution doublement différentiable $F(x)$ tend vers $H_{3,0}(x)$ si, lorsque x tend vers l'infini,

$$\lim_{x \rightarrow \infty} \frac{g'(x)}{g^2(x)} = 0 \quad (4.1)$$

où $f(x)$ désigne la densité de F , $g(x) = f(x)/[1 - F(x)]$ désigne le taux d'échec de F , et g' est la dérivée de g . Nous proposons ci-après une approximation de la constante de winsorisation $R(F, n)$ pour cette classe de distributions.

Proposition 2 Si $F(x)$ est telle que l'équation (4.1) est valide et si, pour les grandes valeurs de x , elle répond aux conditions suivantes:

- i) $xg(x)$ augmente;
- ii) $xg'(x)/g(x)$ est inférieur à une certaine constante positive c ;

alors la constante de winsorisation optimale $R(F, n)$ satisfera

$$R(F, n) =$$

$$F^{-1} \left(1 - \frac{g[F^{-1}(1 - 1/n)] F^{-1}(1 - 1/n) [1 + o(1)]}{n} \right);$$

et $m(F, n) = g(F^{-1}(1 - 1/n)) F^{-1}(1 - 1/n) [1 + o(1)]$. En outre, l'erreur quadratique moyenne de la moyenne winsorisée de Searls sera approximativement égale à:

$$\text{MSE}(\bar{X}_R) \approx \frac{\sigma^2}{n} - \frac{R(F, n)^2}{n^2}.$$

Dans la famille de Weibull, $F^{-1}(1 - t) = [-\log(t)]^\alpha$, $g(x) = x^{1/\alpha - 1}$. Les hypothèses de la proposition 2 sont réalisées et $m(F, n)$, le nombre prévu d'observations winsorisées dans un grand échantillon de Weibull, est donné par $\log(n) [1 + o(1)] / \alpha$ qui tend vers l'infini à mesure que n augmente. La figure 1 donne à penser que la convergence est très lente, en particulier pour les grands coefficients de variation.

Examinons maintenant les distributions dont le maximum de l'échantillon tend vers $H_{1,\alpha}(x)$. Cette classe de distributions a été caractérisée par Gnedenko (1962): le maximum de l'échantillon de F converge vers $H_{1,\alpha}(x)$ si on peut affirmer que

$$1 - F(x) = L(x)/x^\alpha \quad (4.2)$$

où L est une fonction qui tend vers l'infini, $L(x)/L(kx)$ converge vers 1 pour n importe quelle constante k . Notons que pour que F possède un deuxième moment fini, il suffit que $\alpha > 2$ dans (4.2). La distribution de Pareto satisfait à (4.2) avec $\alpha = \gamma$.

Proposition 3 Si F satisfait à (4.2) avec le paramètre $\alpha > 2$, alors, à mesure que n tend vers l'infini, $R(F, n) = F^{-1} [1 - (\alpha - 1)/n] [1 + o(1)]$, c.-à-d., $m(F, n) \approx \alpha - 1$. En outre,

$$\text{MSE}(\bar{X}_R) \approx \frac{\sigma^2}{n} - \frac{\alpha R(F, n)^2}{n^2 (\alpha - 2)}.$$

Pour les distributions qui satisfont à (4.2), un nombre fini d'observations sont en moyenne winsorisées à mesure que l'échantillon tend vers l'infini. Ceci s'observe dans une certaine mesure dans la Figure 1, où les courbes de $m(F, n)$ pour la distribution de Pareto ont $m(F_{2,33}, n) = 1.33$, et $m(F_{2,67}, n) = 1.67$ comme asymptotes. Les propositions 2 et 3 jettent une certaine lumière sur l'estimation de la valeur limite optimale. Lorsque F est inconnu, la valeur qui minimise un estimateur de l'erreur quadratique moyenne de \bar{X}_R est un estimateur possible de $R(F, n)$. On obtient ainsi:

où $F(y) = \sum n_h F_h [\mu_h + n_h y / (n W_h)] / n$. L'équation (3.4) est aisément résolue en utilisant l'algorithme (2.5) proposé dans la section 2 pour le cas de l'échantillon unique. On peut ainsi calculer facilement des approximations simples des valeurs limites optimales de Searls pour les échantillons stratifiés.

Comme la distribution F définie ci-dessus a une espérance de zéro, l'erreur quadratique moyenne de la moyenne stratifiée winsorisée obtenue par la résolution de l'équation (3.3) est égale à :

$$\text{MSE}(\bar{X}_R) = \frac{1}{n} \left(\sigma_F^2 - 2 \int_{-\infty}^R y [1 - F(y)] dy - B(\bar{X}_R)^2 + B(\bar{X}_R)^2 \right) + \left(\frac{1}{n} B(\bar{X}_R)^2 - \sum_{h=1}^L \frac{W_h^2 B^2(\bar{X}_{Rh})}{n_h} \right) \quad (3.5)$$

où σ_F^2 est la variance de F . Il est facile de démontrer que le dernier terme de (3.5) est négatif ou nul ; il est nul lorsque $B(\bar{X}_R) = n W_h B(\bar{X}_{Rh}) / n_h$ pour $h = 1, \dots, L$. La variance de la moyenne stratifiée, $X = \sum W_h \bar{X}_h$, est égale à σ_F^2 / n . Ainsi, une approximation prudente de l'efficacité de \bar{X}_R par rapport à \bar{X} dans un échantillon stratifié aléatoire de taille n tiré de F . Remarquons également que $n[1 - F(R)]$ représente le nombre total prévu de points de données winsorisés dans la strate L .

Le plan de winsorisation optimale obtenu par la résolution de (3.3) possède une forme simple pour beaucoup de règles de répartition. En vertu de la règle de répartition proportionnelle, c.-à-d., $n_h = n W_h$ pour $h = 1, \dots, L$, on obtient $R_h = \mu_h + R$. En vertu de la répartition optimale de Neyman, avec $n_h = n W_h \sigma_h / (\sum W_h \sigma_h)$, où σ_h est l'écart-type de la strate h , on obtient $R_h = \mu_h + \sigma_h R / (\sum W_h \sigma_h)$. Si, en plus, les distributions de X à l'intérieur des strates sont les mêmes sauf pour un changement dans l'emplacement et dans l'échelle, c.-à-d., $F_h = F_0[(x - \mu_h) / \sigma_h] / \sigma_h$ pour une certaine distribution F_0 , alors $F(x) = F_0[x / (\sum W_h \sigma_h)]$. Dans ce cas, les caractéristiques des moyennes winsorisées optimales dans l'échantillonage stratifié et dans l'échantillonage aléatoire simple sont les mêmes. Ainsi, la figure 1 présente le nombre total prévu de points de données winsorisés dans la strate L en fonction de la taille n de l'échantillon total, en vertu de la répartition de Neyman, lorsque F_0 correspond à une des distributions du tableau 1. La figure 2 donne les valeurs correspondantes de l'efficacité.

Les résultats de cette section sont faciles à généraliser à l'échantillonage stratifié sans remise en remplaçant n_h par $n_h / (1 - f_h)$ tout au long des calculs. On dérive aisément les valeurs limites optimales pour l'échantillonage stratifié avec ppt en posant $F_h(x) = \sum p_{hi} I(y_{hi} / (N_h p_{hi}) \leq x)$, où p_{hi} désigne la probabilité de sélection de la i -ième unité de la strate h .

$$\text{MSE}(\bar{X}_R) = \sum_{h=1}^L \frac{W_h^2}{n_h} \left(\sigma_h^2 - 2 \int_{-\infty}^{R_h} (x - \mu_h) [1 - F_h(x)] dx - B^2(\bar{X}_{Rh}) \right) + \left(\sum_{h=1}^L W_h B(\bar{X}_{Rh}) \right)^2 \quad (3.1)$$

où $B(\bar{X}_{Rh})$ représente le biais de \bar{X}_{Rh} en tant que estimateur de μ_h

$$B(\bar{X}_{Rh}) = - \int_{-\infty}^{R_h} [1 - F_h(x)] dx.$$

L'utilisation des dérivées partielles en fonction de R_h , $h = 1, \dots, L$ donne les équations suivantes pour les valeurs optimales :

$$\frac{W_h}{n_h} [R_h - \mu_h - B(\bar{X}_{Rh})] = - \sum_{L=1}^h W_h B(\bar{X}_{Rh}), \quad (3.2)$$

pour $h = 1, \dots, L$.

Il n'existe pas de méthode simple pour résoudre (3.2). On peut cependant obtenir une solution approximative en notant que $B(\bar{X}_{Rh}) / n_h$ est habituellement petit comparativement aux autres éléments de l'équation, pour toutes les valeurs de h . L'élimination de ces éléments conduit à l'équation suivante :

$$\frac{W_h}{n_h} (R_h - \mu_h) = - \sum_{L=1}^h W_h B(\bar{X}_{Rh}), \quad (3.3)$$

pour $h = 1, \dots, L$. Les solutions de (3.3) surestiment légèrement les valeurs optimales qui satisfont à l'équation (3.2) puisqu'elles laissent conclure que les dérivées partielles de (3.1) sont toutes positives et qu'elles représentent des fonctions croissantes de R_h , pour $h = 1, \dots, L$. Ainsi, en résolvant (3.3) pour estimer les valeurs limites, on ne court pas le risque de winsoriser trop de valeurs. L'équation (3.3) signifie que $R_h = \mu_h + n_h R / (n W_h)$, où R est une constante positive. On obtient une équation simple de R en remplaçant la variable $y = n W_h (x - \mu_h) / n_h$ dans les intégrales par $B(\bar{X}_{Rh})$, $h = 1, \dots, L$, où $n = \sum n_h$. On obtient ainsi :

$$- \sum_{h=1}^L W_h B(\bar{X}_{Rh}) = \frac{n}{R}$$

$$\int_{-\infty}^R [1 - F(y)] dy = - B(\bar{X}_R), \quad (3.4)$$

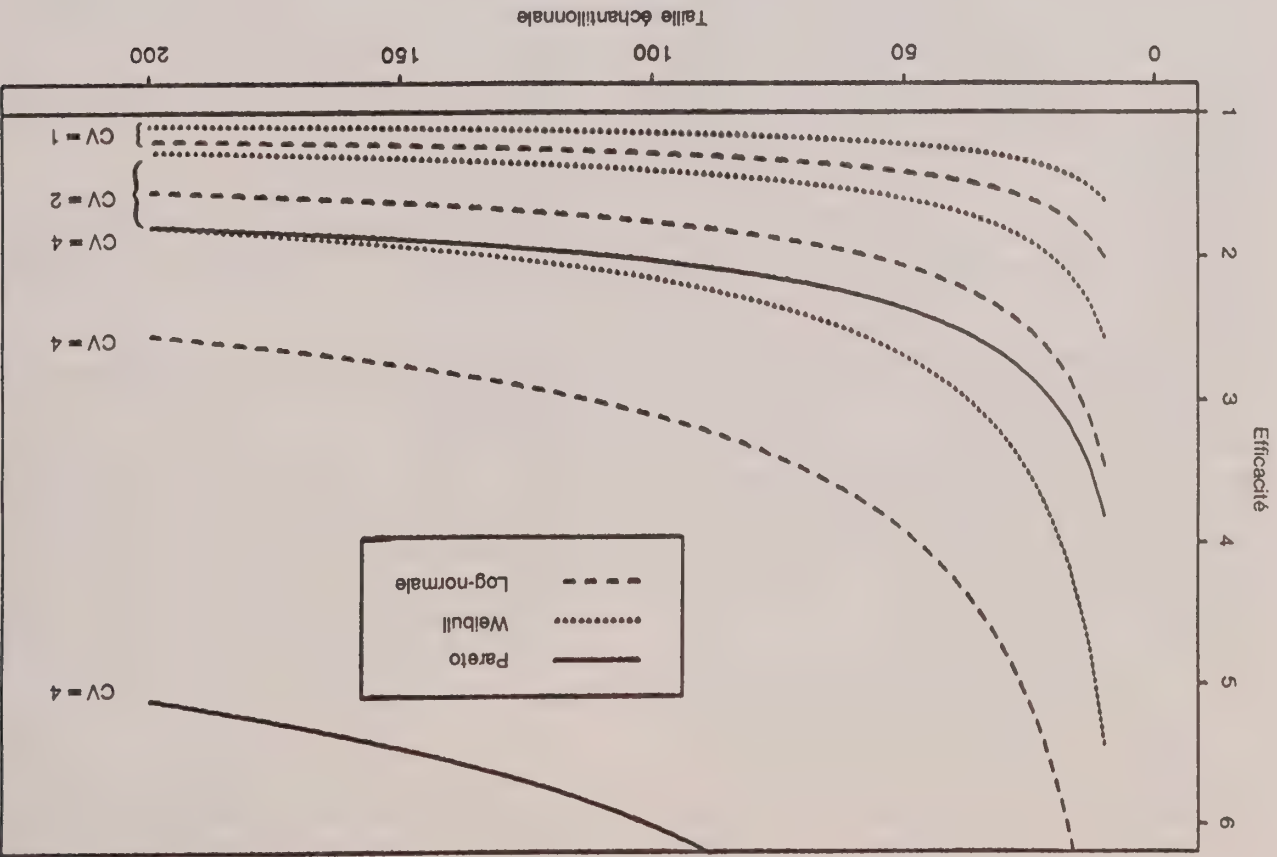


Figure 2. Efficacité de la moyenne winsorisée de Searls.

La figure 1 montre que le nombre prévu d'observations winsorisées $m(F, n)$ diminue avec l'asymétrie croissante de la distribution. Cette observation peut être traduite

en un résultat mathématique rigoureux. À cette fin, on suppose que la variable aléatoire Y est plus asymétrique que la variable aléatoire X , si Y a la même distribution que $\psi(X)$, où $\psi(x)$ est une fonction convexe de x . En vertu de cette définition, X^2 est, comme prévu, plus asymétrique que X . Cette notion de l'asymétrie correspond à l'ordonnancement partiel convexe de van Zwet (Barlow et Proschan 1981). Cette définition de l'asymétrie débouche sur la proposition suivante dont nous fournissons la preuve en annexe, en même temps que celles des propositions 2 et 3.

Proposition 1 Si Y est plus asymétrique que X , alors $m(F_X, n) > m(F_Y, n)$, où F_X et F_Y désignent les distributions de X et de Y respectivement.

Les résultats de la présente section s'appliquent également à l'échantillonnage aléatoire simple sans remise. Pour ce plan d'expérience, l'erreur quadratique moyenne de X_R est donnée par la formule (2.3) en remplaçant n par $n/(1-f)$, où f désigne la fraction d'échantillonnage. L'algorithme (2.5), avec n divisé par $(1-f)$, peut servir

au calcul des valeurs limites optimales pour l'échantillonnage aléatoire simple sans remise.

3. WINSORISATION EN ÉCHANTILLONNAGE STRATIFIÉ

Il existe plusieurs façons d'appliquer la méthode de winsorisation de Searls à l'échantillonnage stratifié. Dans la présente section, chaque strate possède sa propre valeur limite. Désignons par R_h la valeur limite de la strate h . Les valeurs optimales R_1, R_2, \dots, R_L , où L désigne le nombre de strates, sont celles qui minimisent l'erreur quadratique moyenne de $X_R = \sum W_h X_{Rh}$, où $X_{Rh} = \sum \min(X_{hi}, R_h)/n_h$, $W_h = N_h/N$ et N_h désigne la taille de la strate h et $N = \sum N_h$. Nous proposons dans la présente section un algorithme pour la détermination de ces valeurs limites optimales.

Désignons par $F_h(x)$, pour $h = 1, \dots, L$, la distribution de X dans la strate h , et par μ_h et σ_h^2 la moyenne et la variance de F_h respectivement. La détermination de l'erreur quadratique moyenne de X_R , dans un échantillonnage aléatoire stratifié avec remise, se fait de la façon décrite à la section 2. On obtient:

avec $R_0 = 2\mu$ comme valeur de départ, tend doucement vers la solution de (2.4). Pour les distributions discrètes, les calculs sont faits aisément en se rappelant que

$$\int_0^R [1 - F(x)] dx = E[\max(X - R, 0)].$$

Les calculs exacts des points limites optimaux $R(F, n)$ ont été effectués pour les familles d'échantillons de Weibull, log-normale et de Pareto, avec des échantillons dont la taille s oscillait entre 5 et 200. Trois distributions, correspondant aux coefficients de variation (CV) 1, 2, et 4, ont été utilisées pour les deux premières familles alors que pour la famille de Pareto, on utilisait uniquement les coefficients de variation 2 et 4. Le coefficient de variation mesure l'asymétrie de la distribution, les valeurs les plus grandes correspondant à l'asymétrie la plus prononcée. Les valeurs correspondantes des paramètres sont présentées dans le tableau 1.

Tableau 1

Valeurs des paramètres des distributions dont on a évalué les valeurs limites optimales $R(F, n)$

CV	Weibull(α)	Log-normale(ν)	Pareto (γ)
1	1	0.83	-
2	1.84	1.27	2.67
4	2.87	1.68	2.13

On a calculé le point limite optimal pour chaque distribution et pour chaque taille d'échantillon en utilisant l'algorithme (2.5). La figure 1 présente le nombre prévu d'observations winsorisées, $m(F, n) = n\{1 - F[R(F, n)]\}$ en tant que fonction de n , et la figure 2 donne les valeurs correspondantes de l'efficacité. L'efficacité de \bar{X}_R est définie par $\text{Var}(\bar{X})/\text{MSE}(\bar{X}_R)$. Dans la figure 1, le nombre prévu de valeurs winsorisées en vertu du plan optimal s'approche de 1 pour la plupart des distributions asymétriques, même pour les grands échantillons. En faisant l'approximation de ce nombre à l'aide d'une distribution de Poisson avec le paramètre $m(F, n)$, on constate qu'il existe une probabilité non négligeable, avec le plan de winsorisation optimale, qu'aucun des points de données ne soit winsorisé. Cette probabilité augmente avec l'asymétrie de la distribution puisque $m(F, n)$ diminue avec le coefficient de variation CV. Ainsi, pour les échantillons provenant d'une distribution fortement asymétrique, il n'est pas toujours approprié de winsoriser les valeurs les plus grandes. De telles valeurs ne devraient être winsorisées que lorsqu'elles sont très grandes par rapport aux autres. Comme prévu, dans la figure 2, les meilleurs gains en matière d'efficacité sont obtenus lorsque l'asymétrie est prononcée. Ainsi, la stratégie de winsorisation la plus profitable consiste à repérer les deux ou trois valeurs les plus grandes d'un échantillon et d'en réduire l'impact lorsqu'elles sont très grandes par rapport aux autres.

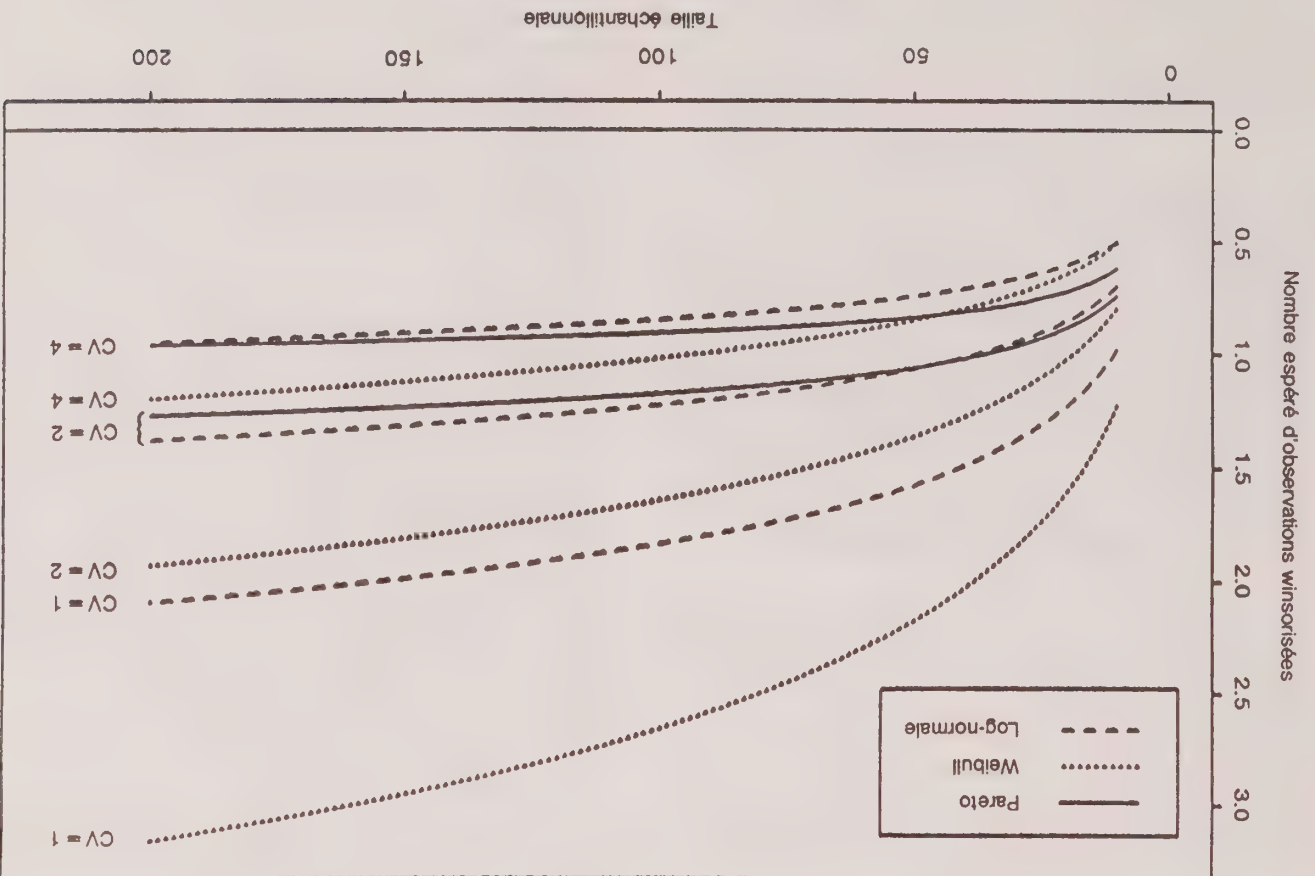


Figure 1. Nombre espéré d'observations winsorisées pour échantillonnage aléatoire simple et stratifié.

à asymétrie positive. On peut en effet choisir entre la famille de Weibull, $F_\alpha(x) = 1 - \exp(-(x/\beta)^{1/\alpha})$ pour $x > 0$; la famille log-normale, $F_p(x) = \Phi(\log(x/\beta)/\sigma)$ pour $x > 0$; et la famille Pareto, $F_p(x) = 1 - (1 + x/\beta)^{-\gamma}$ pour $x > 0$, où β est un paramètre d'échelle positif et α , σ , et γ sont des paramètres de forme positifs. Les distributions asymétriques discrètes proviennent d'échantillonsages d'enquête. Désignons par $\{y_1, \dots, y_N\}$ les valeurs de la variable d'intérêt pour les N unités d'une population à échantillonner. Si on tire un échantillon aléatoire simple avec remise, on peut alors désigner par $F(x) = \sum I(y_i \leq x)/N$ la distribution sous-jacente où $I(\cdot)$ représente la fonction indicatrice. Pour l'échantillonnage avec ppt, c'est-à-dire l'échantillonnage avec remise assorti de probabilités déterminées par $\{p_i, i = 1, \dots, N\}$, on utiliserait $F(x) = \sum p_i I(y_i \leq x)$. L'estimateur standard de y sous échantillonnage avec ppt,

$$\bar{y}_s = \frac{1}{n} \sum_{i=1}^s \frac{y_i}{N p_i}$$

peut alors être considéré comme la moyenne d'un échantillon aléatoire de taille n tiré d'une distribution F . Fuller (1991) donne des exemples de données d'enquête à distribution asymétrique.

Désignons par X_1, X_2, \dots, X_n un échantillon tiré de $F(x)$. Pour l'échantillonnage avec ppt, on obtiendra $X_i = y_i/(N p_i)$ où p_i et y_i désignent la probabilité de sélection et la valeur de la variable y pour la i -ième unité sélectionnée dans l'échantillon. La moyenne de la population μ doit être estimée par une moyenne winsorisée,

$$\bar{X}_R = \frac{1}{n} \sum_{i=1}^n \min(X_i, R) = \bar{X} - \frac{1}{n} \sum_{i=1}^n \max(X_i - R, 0), \quad (2.1)$$

où \bar{X} est la moyenne des valeurs X_i . L'espérance de \bar{X}_R est égale à

$$E(\bar{X}_R) = \mu - \int_{\infty}^R (x - R) dF(x) = \mu - \int_{\infty}^R x dy dF(x).$$

En changeant l'ordre d'intégration dans l'intégrale ci-dessus, on prouve que $E(\bar{X}_R) = \mu + B(\bar{X}_R)$, où

$$B(\bar{X}_R) = - \int_{\infty}^R [1 - F(x)] dx \quad (2.2)$$

représente le biais de la moyenne winsorisée.

Selon la formule (2.1), l'expression de la variance de \bar{X}_R est

$$n \text{Var}(\bar{X}_R) = \sigma^2 - 2 \text{cov}[X_1, \max(X_1 - R, 0)]$$

$$+ \text{Var}[\max(X_1 - R, 0)]$$

où X_1 est la première variable aléatoire de l'échantillon et σ^2 est la variance de $F(x)$. Des manipulations semblables à celles qui aboutissent à la formule (2.2) montrent que

$$E[\max(X_1 - R, 0)^2] = 2 \int_{\infty}^R (x - R)[1 - F(x)] dx,$$

et que

$$E[\max(X_1 - R, 0)X_1] = 2 \int_{\infty}^R (x - R)[1 - F(x)] dx - RB(\bar{X}_R).$$

Ainsi,

$$\text{Var}(\bar{X}_R) =$$

$$\frac{1}{n} \left\{ \sigma^2 - 2 \int_{\infty}^R (x - \mu)[1 - F(x)] dx - B^2(\bar{X}_R) \right\},$$

et

$$\text{MSE}(\bar{X}_R) = \frac{\sigma^2}{2} - \frac{n}{2} \int_{\infty}^R (x - \mu)[1 - F(x)] dx$$

$$+ \frac{n}{n-1} B^2(\bar{X}_R). \quad (2.3)$$

Searls (1966) a démontré que l'erreur quadratique moyenne de \bar{X}_R comporte un minimum unique qui peut être obtenu en égalant sa dérivée, par rapport à R , à 0. On obtient ainsi l'équation suivante pour la constante optimale de winsorisation $R(F, n)$:

$$\frac{R - \mu}{n} - \int_{\infty}^R [1 - F(x)] dx = 0. \quad (2.4)$$

Cette équation équivaut à l'équation (14) de Searls (1966). Dans le reste de ce travail, \bar{X}_R désigne la moyenne winsorisée optimale obtenue par la constante de winsorisation $R(F, n)$ qui résout (2.4). À noter que le point limite optimal $R(F, n)$ est équivariant quant à l'emplacement et à l'échelle, c'est-à-dire que si $G(x) = F[(x - b)/a]$, alors $R(G, n) = aR(F, n) + b$. Il est facile d'établir un algorithme général pour la résolution de (2.4). Notons d'abord qu'en sa qualité de fonction de R , le membre de gauche de l'équation (2.4) est croissant et concave en R puisque sa dérivée, $1/(n - 1) + 1 - F(R)$, est positive et décroissante. En conséquence, l'algorithme de Newton-Raphson (Thisted 1988, 164-167), donné par

$$R_{j+1} = R_j - \frac{(R_j - \mu) - (n - 1) \int_{R_j}^{\infty} [1 - F(x)] dx}{1 + (n - 1)[1 - F(R_j)]}, \quad (2.5)$$

Moyenne winsorisée de Searls pour populations asymétriques

LOUIS-PAUL RIVEST et DANIEL HURTUBISE¹

RÉSUMÉ

Nous examinons l'utilité de la moyenne winsorisée comme estimateur de la moyenne d'une distribution à asymétrie positive. On obtient la moyenne winsorisée en remplaçant toutes les observations supérieures à une valeur limite donnée R par cette même valeur R , avant le calcul de la moyenne. La valeur limite optimale, telle qu'elle est définie par Searls (1966), minimise l'erreur quadratique moyenne (mean square error, ou MSE) de l'estimateur winsorisé. Nous proposons des méthodes d'évaluation de cette valeur limite optimale avec divers plans d'échantillonnage, y compris l'échantillonnage aléatoire simple, l'échantillonnage stratifié et l'échantillonnage avec probabilité proportionnelle à la taille (ppt). Pour la plupart des distributions asymétriques, la stratégie de winsorisation optimale aura généralement pour effet de modifier la valeur d'environ une observation dans l'échantillon. On dérive des approximations directes (qui ne fait pas appel à un processus itératif) de l'efficacité de la moyenne winsorisée de Searls en utilisant la théorie des statistiques d'ordre extrême. Une expérience de Monte Carlo nous sert à comparer divers estimateurs réduisant l'impact des valeurs extrêmes.

MOTS CLÉS: Valeurs aberrantes; domaine d'attraction maximal; erreur quadratique moyenne; échantillonnage aléatoire simple; échantillonnage stratifié.

1. INTRODUCTION

Les échantillons tirés de distributions étalées vers la droite contiennent souvent des valeurs aberrantes beaucoup plus grandes que la plupart des valeurs échantillonnées. On s'efforce habituellement de tenir compte de ces valeurs extrêmes à l'étape de l'élaboration du plan de sondage (Glasser 1962; Hidiroglou 1987). Toutefois, compte tenu des buts multiples de la plupart des sondages, les statisticiens se trouvent souvent aux prises avec des valeurs aberrantes à l'étape de l'estimation. Ces observations nuisent à la stabilité des estimateurs classiques comme la moyenne de l'échantillon. Il paraît donc utile d'étudier des estimateurs de rechange qui réduiront l'incidence des valeurs trop grandes. La winsorisation (Searls 1966) consiste à remplacer les valeurs supérieures à une valeur limite R par cette même valeur R avant le calcul de la moyenne. Searls suggère de choisir un R qui minimise l'erreur quadratique moyenne de la moyenne winsorisée. On peut également choisir un R correspondant à la deuxième valeur en importance dans l'échantillon (Rivest 1994). L'estimateur de Searls s'est avéré le meilleur, parmi toutes les méthodes examinées par Ernst (1980), pour ajuster les valeurs trop grandes. Hicks et Fetter (1993) ont utilisé la méthode de winsorisation de Searls dans le cadre d'un sondage sur l'agriculture. D'autres méthodes ont été proposées pour traiter les valeurs trop grandes pré-sentes dans les échantillons d'enquêtes. Chambers et Kokic (1993) ont étudié les estimateurs dérivés de la théorie des "statistiques robustes" (Huber 1981). Fuller (1991, 1993) a proposé un test préliminaire pour la détection des valeurs extrêmes dans les échantillons pour lesquels ce test est significatif. Lee (1994) présente un bon survol des articles de plus en plus nombreux abordant cette question.

La clé de la mise en oeuvre de la méthode de winsorisation de Searls réside dans la sélection de la valeur limite R . Nous suggérons dans la section 2 un algorithme simple pour le calcul de la valeur limite optimale, pour une population connue soumise à un échantillonnage aléatoire simple ou à un échantillonnage avec ppt. Des calculs répétés de la valeur limite optimale pour plusieurs populations et plusieurs tailles d'échantillons révèlent que dans la plupart des cas, la méthode optimale modifie en moyenne une observation, peu importe la taille de l'échantillon. Dans la section 3 nous reprenons le travail effectué dans la section 2 mais en utilisant cette fois l'échantillonnage stratifié; nous proposons un algorithme simple pour le calcul des valeurs limites dans chaque strate. La règle de la winsorisation d'une observation en moyenne, sans égard à la taille de l'échantillon, s'applique également aux échantillons stratifiés. Dans les sections 4 et 5, nous calculons l'efficacité, en ce qui concerne la moyenne de l'échantillon, de divers estimateurs winsorisés. Dans la section 4, nous dérivons des approximations de l'efficacité de l'estimateur de Searls pour les grands échantillons au moyen de la théorie des statistiques d'ordre extrême. Dans la section 5, nous comparons l'utilité des estimateurs pour la réduction de l'incidence des valeurs extrêmes à l'aide d'une étude de Monte Carlo.

2. PROPRIÉTÉS D'ÉCHANTILLONNAGE DE LA MOYENNE WINSORISÉE

Nous examinons dans la présente section les moyennes winsorisées pour des données tirées d'une distribution discrète ou continue. Plusieurs familles de distributions continues peuvent nous servir de modèles de distributions

¹ Louis-Paul Rivest et Daniel Hurtubise, Département de mathématiques et de statistique, Université Laval, Cité Universitaire (Québec) Canada, G1K 7P4.

Comportement ancien	
Paiement des premiers honoraires:	"Vers quel âge environ avez-vous consulté un dentiste pour la première fois et payé, vous ou vos parents, ses honoraires?"
Consultations:	"Avez-vous vu un dentiste au cours des 12 derniers mois?"
Année de la dernière consultation:	"Quelle année avez-vous consulté un dentiste pour la dernière fois?"
Coût l'an dernier:	"Combien avez-vous dépensé pour des soins dentaires au cours des 12 derniers mois?"

BIBLIOGRAPHIE

ANDERSON, D.A. (1985). Variance component models with binary response; interviewer variability. *Journal of the Royal Statistical Society, Series B*, 47, 203-210.

BIEMER, P.B., et STOKES, S.L. (1985). Optimal design of interviewer variance experiments in complex surveys. *Journal of the American Statistical Association*, 80, 158-166.

BEBBINGTON, A.C., et SMITH, T.M.F. (1977). The effect of survey design on multivariate analysis, *The Analysis of Survey Data*, (C.A. O'Muircheartaigh et C. Payne, Eds.), 175-192. New York: John Wiley.

CHOI, I.C., et COMSTOCK, G.W. (1975). Interviewer effects on responses to a questionnaire relating to mood. *American Journal of Epidemiology*, 101, 84-92.

COLLINS, M., et BUTCHER, B. (1992). Interviewer and clustering effects in an attitude survey. *Journal of the Market Research Society*, 25, 39-58.

CUTRESS, T.W., HUNTER, P.B., DAVIS, P.B., BECK, D.J., et CROXSON, L.J. (1979). *Adult Oral Health and Attitudes to Dentistry in New Zealand*, Medical Research Council, Wellington.

DIJKSTRA, W. (Ed.) (1982). *Response Behaviour in the Survey Interview*. New York: Academic Press.

FEATHER, J. (1973). *A Study of Interviewer Variance*, (WHO/ICS-MCU Saskatchewan Study Area Reports Series 2, No. 3). Department of Social and Preventive Medicine, University of Saskatchewan, Saskatoon.

FELLEGI, I.P. (1974). An improved method of estimating correlated response variance. *Journal of the American Statistical Association*, 69, 496-501.

GROVES, R. (1989). *Survey Errors and Survey Costs*. New York: John Wiley.

GROVES, R., et FULLTZ, N.H. (1985). Gender effects among telephone interviewers in a survey of economic attitudes. *Sociological Methods and Research*, 14, 31-52.

HOLT, D., SMITH, T.M.F., et WINTER, P.D. (1980). Regression analysis of data from complex surveys. *Journal of the Royal Statistical Society*, 143, 474-487.

HOX, J.J. (1994). Hierarchical regression models for interviewer and respondent effects. *Sociological Methods and Research*, 22, 300-318.

KISH, L. (1965). *Survey Sampling*. New York: John Wiley.

KISH, L., et FRANKEL, M.R. (1974). Inferences from complex samples. *Journal of the Royal Statistical Society, Series B*, 36, 1-37.

KISH, L. (1987). *Statistical Design for Research*. New York: John Wiley.

LESSLER, J.T., et KALSBEEK, W.D. (1992). *Nonsampling Errors in Surveys*. New York: John Wiley.

O'MUIRCHEARTAIGH, C.A. (1976). Response errors in an attitudinal sample survey. *Quality and Quantity*, 10, 97-115.

O'MUIRCHEARTAIGH, C.A., et PAYNE, C. (Eds.) (1977). *The Analysis of Survey Data*, (Volume 2: Model Fitting). New York: John Wiley.

O'MUIRCHEARTAIGH, C.A., et WIGGINS, R.D. (1981). The impact of interviewer variability in an epidemiological survey. *Psychological Medicine*, 11, 817-824.

PANNENKOEK, J. (1988). Interviewer variance in a telephone survey. *Journal of Official Statistics*, 4, 375-384.

PRASAD, N.G., et RAO, J.N.K. (1990). The estimation of mean squared error of small area estimators. *Journal of the American Statistical Association*, 85, 163-171.

RAO, J.N.K., et THOMAS, D.R. (1988). The analysis of cross-classified data from complex sample surveys. *Sociological Methodology*, 18, 213-269.

SKINNER, C.J., HOLT, D., et SMITH, T.M.F. (Eds.) (1989). *Analysis of Complex Surveys*. New York: John Wiley.

VERMA, V., SCOTT, C., et O'MUIRCHEARTAIGH, C.A. (1980). Sample designs and sampling errors for the World Fertility Survey. *Journal of the Royal Statistical Society, Series A*, 143, 431-473.

dans un plan d'échantillonnage à degrés multiples, en fonction des hypothèses de départ du modèle de réponses

Le modèle simple sert à déterminer l'incidence relative des effets du groupement de l'échantillon et de l'intervieweur sur l'estimation des moyennes et des proportions. Les résultats confirment plusieurs constatations déjà fort bien documentées: les corrélations intra-catégorie sont habituellement plus faibles pour l'intervieweur que pour le groupement; les corrélations intra-catégorie des groupements varient dans le sens prévu selon le type de question; l'effet global du plan d'échantillonnage se situe entre 2 et 3,5 pour les mêmes types de question; la variabilité de l'intervieweur contribue de manière importante à l'inflation observée et résulte peut-être du facteur d'amplification attribuable à la lourde charge de travail; enfin, l'incidence des effets dus à l'intervieweur varie dans le sens prévu selon le type de question.

En deuxième lieu, on s'est servi du modèle élargi pour analyser les effets dus au groupement et à l'intervieweur sur l'estimation de l'écart entre les moyennes et les proportions pour différents domaines, soit deux jeux de comparaisons selon le sexe et la race. L'effet sur l'écart entre les moyennes des domaines est plus faible mais reste significatif pour plusieurs paramètres, notamment les comparaisons selon la race, ce qui donne à penser que l'intervieweur n'a pas la même incidence sur chaque domaine. L'effet est néanmoins suffisamment important pour qu'on s'en inquiète lors de l'élaboration d'études analogues. En général, les effets dus à l'intervieweur ont deux ou trois fois plus d'incidence sur les comparaisons selon la race que sur celles se rapportant au sexe.

Les carences fondamentales de l'analyse signifient que les résultats ont plus une valeur d'indication que de preuve irrefutable. Ils révèlent cependant que le recours à un groupe restreint d'intervieweurs peut avoir de sérieuses conséquences, même lorsqu'on s'intéresse plus aux différences entre domaines qu'aux moyennes ou aux proportions simples. Cette constatation se démarque sans aucun doute des convictions bien ancrées dans certains milieux, comme celui de la santé, et indique que des recherches empiriques considérablement plus poussées s'avèreraient tout à fait justifiées.

Partant des hypothèses du modèle simple de réponses corrélées, on conclut qu'il est possible d'atténuer l'incidence de la variance de l'intervieweur en augmentant le nombre d'intervieweurs, donc en réduisant la charge de travail de chacun d'eux. Bien sûr, il pourrait s'ensuivre une perte au niveau de la qualité des entrevues si on est contraint, pour la même raison, à rogner sur la formation et les méthodes de surveillance. Dans ce cas, on portera une grande attention à la façon dont sont formulées les questions et instructions destinées aux intervieweurs. Pour ce qui est de la version élargie du modèle cependant, une telle stratégie ne suffira sans doute pas. Si l'un des principaux objectifs de l'étude est la comparaison de différents groupes, il importera de veiller à ce que les intervieweurs traitent les groupes concernés de la manière la plus uniforme qu'il soit. L'enquête devra absolument être conçue

REMERCIEMENTS

Le projet a bénéficié de l'aide du Medical Research Council of New Zealand. Joanna Broad s'est occupée dans une large mesure des calculs détaillés de l'analyse.

ANNEXE

Questionnaire

Attitudes

Dentistes 1: "Les dentistes s'intéressent plus à leurs patients qu'à gagner de l'argent."

Dentistes 2: "Les dentistes recommandent beaucoup plus de choses que le strict minimum."

Dentier: "Un dentier est aussi bon (meilleur) que les dents naturelles."

Fluoration: "Que pensez-vous de la fluoration des réserves d'eau publiques?"

Consultations: "Croyez-vous qu'une personne devrait aller chez son dentiste uniquement lorsqu'elle éprouve un problème ou devrait aussi le consulter quand tout va bien?"

Etat des dents: "Si vous aviez rendez-vous chez le dentiste demain, celui-ci trouverait-il quelque chose à redire au sujet de vos dents?"

Gencives: "Si vous aviez rendez-vous chez le dentiste demain, celui-ci trouverait-il quelque chose à redire au sujet de vos gencives?"

Comportement récent

"Hier, avez-vous utilisé un révélateur/un

rinçage-bouche/de la soie dentaire/un

cure-dent?

- vous êtes-vous rincé la bouche après

- avoir mangé?

- vous êtes-vous brossé les dents?"

"Avez-vous acheté des sucreries ou du chocolat la semaine dernière?"

entre les domaines concernés. En supposant que la charge de travail des intervieweurs est répartie de façon uniforme entre les domaines, v_j est égal à zéro quand l'effet dû à l'intervieweur est identique dans les deux cas, car il y a annulation lors de la comparaison. Si les effets dans les deux domaines sont faiblement corrélés, par contre, v_j a tendance à avoir une valeur beaucoup plus élevée, valeur qui, dans les cas extrêmes, pourrait être égale à la charge de travail moyenne. Dans l'enquête qui nous intéresse, les valeurs de v_j se situent entre 0 et 50 pour le sexe et la race. Les effets dus à l'intervieweur spécifiques au domaine peuvent donc avoir une incidence passablement importante sur l'effet du plan d'échantillonnage. Les mêmes remarques s'appliquent à l'incidence du groupement sur la comparaison; si les effets sont les mêmes dans les deux domaines, ils s'annulent dans une large mesure et l'incidence nette est minimale, mais cette dernière peut être relativement importante quand l'effet du groupement est spécifique au domaine.

Tableau 2

Effets de l'intervieweur et du groupement sur l'écart entre les domaines

Question	Selon le sexe				Selon la race			
	p_j	p_c	D_2	%Int	p_j	p_c	D_2	%Int

Attitudes:								
Paramètres socio-démographiques:								
Race	.008	.183	1.11	0	—	—	—	—
Revenu familial	.004	.095	1.37	24	.004	.099	1.95	40
Etat civil	.000	.059	1.17	0	.011	.060	1.69	38
Situation de l'emploi	.014	.067	1.42	71	.022	.116	2.09	25
Âge	.007	.052	1.06	0	.006	.093	1.87	24
Sexe	—	—	—	—	.006	.011	1.09	44
Moyenne	.007	.091	1.23	19	.010	.076	1.74	34
Comportement récent:								
Brosser les dents	.025	.060	1.62	65	.019	.019	1.68	88
Rincer la bouche	.007	.029	1.28	64	.004	.023	1.20	45
Rince-bouche	.000	.057	1.20	0	.027	.105	2.69	75
Soie dentaire	.003	.021	1.06	0	.015	.036	1.37	32
Cure-dents	.006	.010	1.03	0	.006	.046	1.48	63
Suceries/chocolat	.012	.009	1.02	0	.013	.022	1.31	48
Dentifrice au fluorure	.010	.007	1.11	100	.007	.000	1.02	100
Moyenne	.009	.028	1.19	33	.013	.036	1.36	64
Comportement ancien:								
Âge: première consultation payée	.003	.033	1.10	0	.029	.141	2.92	71
Consultation	.005	.035	1.04	0	.020	.018	1.26	50
Année de la dernière consultation	.004	.012	1.20	75	.016	.003	1.83	12
Coût l'an dernier	.007	.021	1.01	0	.076	.117	2.09	42
Moyenne	.005	.025	1.09	19	.035	.070	2.03	44

Le tableau 2 donne les valeurs de p_j et p_c pour les comparaisons selon le sexe et selon la race, ainsi que l'effet global D_2 du plan d'échantillonnage et la partie de cet effet attribuable à la variabilité de l'intervieweur. Notons que la question relative à l'usage d'un révélateur a été supprimée au tableau 2, principalement parce que très peu de répondants se servaient d'un tel produit ou en connaissaient l'existence, si bien que la taille de l'échantillon utile dans ce cas est faible et que les résultats ne sont pas significatifs.

L'intervieweur et le groupement ont relativement peu d'incidence sur les comparaisons selon le sexe et les effets du plan d'échantillonnage dépassent à peine l'unité, bien que les valeurs estimées de p_j et p_c augmentent légèrement, après correction pour cette variable. On ne relève d'effet lié au sexe significatif que pour trois questions, soit l'état des dents et l'usage d'une brosse à dents – pour lesquelles il pourrait exister un biais particulier dû à l'acceptabilité sociale – et la situation de l'emploi – qui présente des connotations fort distinctes pour les hommes et les femmes. Remarquons que l'effet dû à l'intervieweur domine dans chacune des trois comparaisons.

L'incidence des mêmes facteurs sur les comparaisons selon la race est beaucoup plus élevée, l'effet global du plan d'échantillonnage s'établissant en moyenne à 1.7. On peut en conclure que la race des répondants entraîne des effets dus à l'intervieweur et au groupement significatifs qui ne s'annulent pas. D'importants effets dus à l'intervieweur sont bien corrélés à la race pour deux questions hypothétiques sur les attitudes (consultations et fluoration), pour une question concernant le comportement récent ainsi que pour l'âge au moment où sont pour la première fois payés les honoraires du dentiste. Ces résultats sont plausibles; tous les intervieweurs étaient européens d'origine, ce qui pourrait avoir donné lieu à une variation systématique de leur interaction avec les répondants d'une autre ethnie. Ici encore, les effets du groupement sont plus marqués pour les paramètres sociodémographiques. Non seulement les effets du plan d'échantillonnage sont-ils en moyenne plus élevés que ceux notés pour les comparaisons selon le sexe, mais les effets dus à l'intervieweur sont aussi deux ou trois fois plus élevés en général pour l'écart entre les groupes ethniques.

Un examinateur objectif a fort justement souligné qu'étant donné la manière dont les intervieweurs sont déployés (ils travaillaient principalement en équipe dans diverses parties de la Nouvelle-Zélande), il se pourrait fort bien que les effets dus à l'intervieweur soient exagérés parce qu'ils se confondent avec les effets dus à la région. Les différences entre les DL sont si ténues cependant qu'une telle exagération serait minimale, mais on ne peut effectivement pas écarter cette possibilité avec un modèle ainsi constitué.

5. DISCUSSION

Le présent article se sert des données empiriques issues d'une enquête typique sur la santé pour évaluer l'incidence de la variabilité de l'intervieweur sur la variance de l'erreur

Tableau 1
Effets du groupement et de l'intervieur sur les moyennes et les proportions

Question	ρ_I	ρ_C	D_0	% Int
Attitudes:				
Dentistes 1	.014	.014	4.61	91
Consultations	.008	.028	3.42	74
Dents naturelles	.008	.027	3.52	76
Etat des dents	.007	.015	2.97	84
Dentier	.005	.015	2.67	80
Dentistes 2	.004	.033	2.77	57
Gencives	.003	.010	1.96	77
Fluoruration	.001	.016	1.66	49
Moyenne				
	.006	.020	2.95	73
Paramètres socio-démographiques:				
Situation de l'emploi	.010	.055	4.20	65
Race	.009	.172	6.87	33
Âge	.004	.042	5.98	52
Revenu familial	.002	.092	3.29	15
Etat civil	.000	.058	2.34	0
Sexe	.000	.005	1.12	0
Moyenne				
	.004	.071	3.47	28
Comportement récent:				
Brosser les dents	.019	.025	6.16	8
Sucreries/chocolat	.011	.003	3.75	98
Dentifrice au fluorure	.008	.000	3.04	100
Cure-dents	.006	.006	2.66	92
Rincer la bouche	.004	.024	2.43	62
Soie dentaire	.001	.018	1.60	43
Révélateur	.000	.027	1.49	0
Rince-bouche	.000	.018	1.42	0
Moyenne				
	.006	.012	2.82	60
Comportement ancien:				
Âge: première consultation payée	.004	.029	2.34	57
Consultation	.004	.029	2.51	57
Coût l'an dernier	.002	.000	1.19	100
Année de la dernière consultation	.000	.014	1.15	0
Moyenne				
	.003	.018	1.80	54

questions concernant les attitudes produisent des valeurs de ρ_I supérieures à la moyenne, tout comme certains comportements peuvent être sensibles à un biais important résultant de la "désirabilité sociale", par exemple le fait de se brosser les dents ou d'acheter des bonbons et du chocolat. L'origine ethnique et la situation de l'emploi donnent également des valeurs relativement élevées. Les résultats se rapprochent de ceux obtenus lors d'études antérieures, bien qu'ils se trouvent à l'extrême inférieure de la fourchette des valeurs rapportées ailleurs (Feather 1973; Kish 1962; O'Muircheartaigh 1977; O'Muircheartaigh et Wiggins 1981). On trouvera une étude complète de la question au chapitre 8 de l'ouvrage de Groves (1989). Il se peut que ces résultats viennent en partie de la formation et de la surveillance intensives des intervieweurs, associées au volet de l'enquête effectué sur

le terrain. Une "épuración" (correction et vérification) rigoureuse des données après le travail sur le terrain, avant analyse, serait une autre explication. Néanmoins, il se pourrait qu'on doive simplement la situation à l'atténuation qui résulte de l'application du modèle (1) aux proportions mentionnées précédemment.

L'incidence de la variabilité de l'intervieur sur l'effet global du plan d'échantillonnage, qui combine les effets dus à l'intervieur et au groupement, a sans doute plus d'importance que les caractéristiques et les valeurs de ρ_I . C'est ce qu'on remarque à la troisième colonne du tableau 1 (D_0), la dernière colonne (int. %) représentant la contribution proportionnelle de la variabilité de l'intervieur à la valeur globale de D_0 . Le plan d'échantillonnage a des effets sensibles puisque la valeur de ces derniers dépasse deux dans tous les cas, à quelques exceptions près. On le doit au groupement et à l'incidence de la lourde charge de travail des intervieweurs, caractéristique à l'étude puisque, selon l'équation (3), la variance augmentée d'un facteur de $1 + (n_I - 1)\rho_I$, où n_I représente la charge de travail moyenne pondérée des intervieweurs. La variabilité de l'intervieur contribue différemment aux effets du plan d'échantillonnage. Dans le cas des variables socio-démographiques, sa contribution se situe en moyenne un peu au-dessus de la moitié de la contribution attribuable au groupement, alors que pour les questions se rapportant aux attitudes, la contribution de l'intervieur aux effets du plan d'échantillonnage représente le triple de celle due au groupement. La contribution des deux autres catégories de questions se situe entre ces deux extrêmes.

Les résultats du tableau 1 confirment l'influence de la charge de travail de l'intervieur sur la variance des estimations de l'échantillon, en raison du facteur d'amplification. En deux mots, un élément associé à l'intervieur qui présente une très petite corrélation dans la catégorie peut avoir des effets majeurs si la charge de travail de l'intervieur est suffisamment importante. Dans l'enquête à l'étude, la logistique du déploiement des intervieweurs et les exigences relatives à un contrôle de la qualité soutenu semblent précher en faveur de petites équipes, pratique apparemment courante pour bon nombre d'enquêtes sur le terrain dans le secteur de la santé (lire, par exemple, Choi et Comstock 1975). Bref, le nombre d'entrées se situe en moyenne au-dessus de 250. On se rend tout de suite compte du coût d'une telle stratégie en examinant le tableau 1; de très petites variations entre intervieweurs réduisent considérablement la précision des estimations de l'échantillon.

Passons maintenant à l'objet principal de notre analyse, soit l'incidence de la variabilité de l'intervieur sur l'écart entre les moyennes ou les proportions des domaines. Nous l'avons évaluée pour deux jeux de comparaisons, le premier selon le sexe (masculin/féminin) et le second selon la race (européenne/non européenne). Ainsi qu'on a pu le constater dans la discussion suivant l'équation (1), la contribution $1 + (n_I - 1)\rho_I$, à D_0 attribuable à la variation entre intervieweurs dépend de la mesure dans laquelle l'effet dû à l'intervieur est constant dans les deux domaines et de la manière dont la charge de travail individuelle est répartie

binaires (lire Anderson et Aitken 1985 et Pannack 1988 notamment). Les effets du plan d'échantillonnage sur les proportions que nous avons estimés devaient donc être considérés comme la limite inférieure de la fourchette. Les modèles sont ajustés au moyen du module PROC GLM du SAS. Les effets du groupement sont bien expliqués dans la documentation existante (Kish 1965; Kish et Frankel 1970; 1974). En général, leur ampleur dépend du genre et du nombre d'unités sélectionnées et varie sans doute avec tel ou tel paramètre sociodémographique. Dans l'enquête qui nous intéresse, on prévoyait des effets relativement élevés puisque les îlots de recensement servant d'unités d'échantillonnage devaient se caractériser par une assez bonne homogénéité interne. Face à une telle concentration des caractéristiques de la population, on a supposé que les paramètres démographiques et les éléments connexes donneraient les valeurs les plus importantes de p_c . On s'attendait à ce que p_i ait une valeur plus faible en raison de l'importante formation des intervieweurs. La documentation existante suggère que ces effets varieront aussi vraisemblablement avec le genre de questions, par exemple celles sur le comportement, les questions d'approfondissement, celles à choix obligatoire ainsi que les questions mal formulées ou ambiguës, particulièrement sensibles à la variabilité de l'intervieweur (Feather 1973, Groves 1989). Les deux premières colonnes du tableau 1 présentent la mesure estimée des coefficients de corrélation intra-intervieweur et intra-groupe pour diverses questions réparties entre quatre catégories (paramètres sociodémographiques, attitudes, comportement récent et comportement ancien). Une catégorisation de ce genre crée des groupes naturels susceptibles de faire ressortir bon nombre des effets dus à l'intervieweur. Dans chaque groupe, les éléments sont énumérés par ordre de grandeur de la corrélation intra-intervieweur. L'annexe décrit les questions posées (outre celles sur les paramètres sociodémographiques, qui s'expliquent d'elles-mêmes).

Comme on pouvait s'y attendre, les variables socio-démographiques (sauf le sexe) présentent les valeurs les plus élevées pour la corrélation intra-groupe. p_c a une valeur moyenne de 0.07 (0.08 lorsqu'on exclut le sexe). La valeur moyenne de p_c pour les trois autres catégories ne dépasse pas 0.02. Il se peut que certains aspects que la logique associerait étroitement à la situation sociale – par exemple la fréquence des consultations, le paiement des consultations, le brossage des dents et quelques déclarations relatives aux attitudes – présentent une valeur de p_c supérieure à la moyenne. En général cependant, la valeur obtenue reste dans la fourchette signalée par d'autres auteurs (lire, par exemple, Kish 1965, p. 581 pour une série d'enquêtes auprès des consommateurs; Bebbington et Smith 1977 et Verma et coll. 1984 pour les études nationales de l'Enquête mondiale sur la fécondité).

Les valeurs estimées de p_i correspondantes apparaissent à la première colonne du tableau 1. Dans l'ensemble, les valeurs obtenues sont nettement plus faibles que celles relevées pour les effets du groupement (habituellement moins de la moitié, et dans certains cas le dixième de p_c pour les éléments correspondants). Comme prévu, certaines

autres éléments devraient être attribués au hasard. On le fait rarement avec les enquêtes de grande envergure, car il s'agit d'une méthode onéreuse et compliquée. Il s'ensuit qu'on peut difficilement utiliser les résultats de telles enquêtes pour la recherche méthodologique, car il pourrait y avoir confusion entre les paramètres de l'intervieweur et ceux du répondant. L'analyse à degrés multiples comme celle décrite plus haut, y remédie dans une certaine mesure. Les variables pertinentes des répondants, si elles sont connues, peuvent être intégrées au modèle de régression, ce qui permet d'uniformiser les intervieweurs par des moyens statistiques. . . Une telle approche a cependant ses limites. En effet, cette dernière s'appuie sur une méthode de contrôle statistique plutôt qu'expérimentale. Elle repose aussi sur l'hypothèse que le modèle intègre constamment toutes les covariables pertinentes. Sans randomisation, nul ne peut conclure que l'influence de toutes les variables confonduelles a été éliminée⁴ [Traduction]. Dans le cas qui nous intéresse, le déploiement des intervieweurs permet d'identifier formellement tous les éléments de la variance, pourvu que le modèle soit crédible et que l'on accepte que l'attribution du travail aux intervieweurs n'a aucun lien avec les effets du groupement. L'absence de véritable randomisation signifie que les variations dans le genre de réponses obtenues par les intervieweurs pourraient toujours résulter de la répartition de la charge de travail plutôt que du style de l'intervieweur. De toute évidence, les résultats empiriques ne donnent qu'une indication, bref font ressortir des possibilités qu'on devra approfondir au moyen d'études expressément conçues pour cela.

Même en oubliant le manque de randomisation du déploiement des intervieweurs, on se rend compte que le plan d'échantillonnage est beaucoup plus complexe que celui supposé dans la partie théorique qui précède. En effet, l'enquête prévoit trois degrés d'échantillonnage et une stratification régionale des unités à la première étape. Pour l'analyse complète, nous avons ajusté le modèle d'avantage en y intégrant des effets constants pour la stratification, un modèle hiérarchique à effet aléatoire pour les trois niveaux d'échantillonnage et tous les termes d'interaction du deuxième degré. Malgré cela, les DL correspondent aux unités de la première étape étaient si faibles que l'écart entre les strates et la variance entre les DL sont négligeables pour toutes les variables qui reviennent dans l'analyse subséquente. La variance entre USE domine donc et, à toute fin pratique, on peut considérer que l'enquête repose sur un plan d'échantillonnage à deux degrés où les îlots de recensement (regroupés s'il y a lieu) font office d'USE. Nous n'avons pas tenu compte des autres éléments dans les résultats examinés plus bas.

4. RÉSULTATS

m_I ont des valeurs analogues et l'effet de l'intervieur sur l'écart est comparable à celui d'une moyenne simple. En règle générale, les intervieweurs interrogent des personnes des deux domaines et m_I est assez faible, ce qui semble corroborer dans une certaine mesure l'hypothèse que la variabilité de l'intervieur n'exerce qu'une faible influence sur l'écart estimatif entre deux domaines. Les mêmes remarques s'appliquent aux effets du groupement. L'analyse qui précède repose sur l'hypothèse selon laquelle les effets de l'intervieur et du groupement, a_i et b_i , sont identiques pour les deux domaines. On peut facilement imaginer des situations où pareille supposition ne tiendrait pas. Certains intervieweurs, par exemple, réagissent très différemment avec un homme ou une femme, ou les membres de différents groupes ethniques. Le modèle qui suit tient compte de telles interactions éventuelles :

$$Y_{ipd}^{ipd} = \mu_{(d)} + a_{(d)} + b_{(d)} + e_{ipd}^{ipd} \quad (9)$$

où $a_{(d)}$ et $a_{(b)}$ (respectivement $b_{(d)}$ et $b_{(b)}$) sont désor-mais censés être des variables aléatoires corrélées selon $r_I(r_C)$. Le modèle simpliste (4) illustre le cas particulier où les variances des effets sont égales et où la valeur de r_I et de r_C est un. D'un autre côté, si les effets de l'intervieur (du groupement) varient sensiblement d'un domaine à l'autre, $r_I(r_C)$ sera faible (voire aura une valeur négative dans les cas extrêmes). Dans le reste de notre article, nous supposons pour plus de simplicité que les variances de $a_{(d)}$ et $a_{(b)}$ (soit $b_{(d)}$ et $b_{(b)}$ respectivement) sont identiques. Pareille supposition peut ou non être raisonnable dans la pratique, mais la simplification qu'elle permet nous aidera à nous concentrer sur les aspects essentiels. Le modèle de base se rapproche du cas général, mais ses éléments sont légèrement plus complexes. Selon l'équation (9), la variance prévue de $Y_{(a)} - Y_{(b)}$ est :

$$V(Y_{(a)} - Y_{(b)}) = (v_I \sigma_I^2 + v_C \sigma_C^2 + \sigma^2) \left(\frac{1}{1} + \frac{n}{1} \right)$$

$$v_I = \sum_i (d_{(a)}^i)^2 - 2r_I d_{(a)}^i d_{(b)}^i + (d_{(b)}^i)^2 \left/ \left(\frac{1}{1} + \frac{n}{1} \right) \right.$$

Le facteur d'inflation de la variance en vertu de ce nouveau modèle est :

$$D_2 = 1 + (v_I - 1)\rho_I + (v_C - 1)\rho_C. \quad (11)$$

Il s'agit d'une fonction décroissante de r_I , c'est-à-dire de la corrélation entre les effets de l'intervieur pour les deux domaines; plus faible est la corrélation, plus importante sera la hausse de la variance. Quand $r_I = 1$, v_I est égal à m_I et l'effet de l'intervieur est négligeable, pourvu que tous les intervieweurs interrogent un nombre

raisonnablement équilibré de personnes dans les deux domaines. Si r_I est faible cependant (signe qu'il existe une forte interaction entre les intervieweurs et les domaines), v_I a le même ordre de grandeur que m_I et la variabilité de l'intervieur peut influencer sensiblement sur la variance de l'écart entre les domaines.

Dans la pratique, les résultats se situent entre les deux extrêmes et leur incidence probable devra être établie de manière empirique. Dans la prochaine partie, nous commentons donc à recueillir des données pratiques en nous servant des réponses à diverses questions posées lors d'une simple enquête sur la santé caractéristique au type de recherches pour lesquelles les comparaisons de domaines présentent une grande importance (même si l'enquête en question n'avait pas été spécifiquement conçue pour l'usage auquel nous la destinons!).

3. EXEMPLE

Notre exemple repose sur les données issues d'une enquête concernant l'hygiène buccale, les attitudes et les habitudes des Néo-zélandais adultes. On trouvera une description détaillée de l'enquête ailleurs (Cutress et coll. 1979). Les aspects de l'étude qui nous intéressent le plus dans le cadre de la recherche actuelle sont le plan d'échantillonnage et le déploiement des intervieweurs.

L'échantillon a été constitué au moyen d'un plan d'échantillonnage stratifié à degrés multiples. Le pays a été divisé en 256 districts locaux (DL) et on a tiré un échantillon géographiquement stratifié de 68 DL des 256 précédents, en fonction de probabilités de sélection proportionnelles à la taille (PPT) de la population à la première étape, la taille de la population correspondant au nombre estimé de personnes de 15 ans et plus. Chaque DL échantillonné a été divisé en unités secondaires d'échantillonnage (USE) constituées des îlots de recensement existants, regroupés afin de donner un échantillon minimal de cinquante personnes le cas échéant. Deux USE ont ensuite été sélectionnées par PPT pour chaque DL échantillonné à la deuxième étape. Enfin, on a systématiquement retenu 28 adultes dans chaque USE. De cette façon, on a uniformisé la probabilité de sélection finale pour tous les adultes, si bien que l'échantillon est (pratiquement) autopondéré. L'aspect principal de l'enquête concerne le déploiement des intervieweurs. On a recouru aux services de treize intervieweurs, soit au moins trois pour chaque USE. Au moins 10% des personnes que devait interroger les intervieweurs venaient de la même région (Auckland). Idéalement, l'attribution du travail aux intervieweurs devrait faire partie du plan d'échantillonnage, comme le recommandent Fellegi (1974) ou Biemer et Stokes (1985). Malheureusement, l'enquête n'avait pas pour but d'estimer la variance due à l'intervieur et les répondants ont été répartis de façon anarchique plutôt qu'au moyen d'une véritable méthode de randomisation.

Il s'agit d'une situation assez caractéristique aux études importantes. La citation de Hox (1994) que voici la résume bien : "Idéalement, dans les enquêtes où interviennent des

intervieweurs. Voici un modèle de réponses corrélées simple qui convient aux observations effectuées dans le cadre d'une telle enquête:

$$(1) \quad Y_{ipr} = \mu + a_i + b_p + e_{ipr},$$

où i désigne l'intervieweur, p l'unité primaire d'échantillonage (UPÉ) et r , le répondant. La moyenne μ est constante et on suppose que les autres éléments (a_i , b_p et e_{ipr}) sont des variables aléatoires indépendantes dont la variance respective est σ_i^2 , σ_p^2 et σ^2 . Les modèles de ce genre sont abondamment utilisés dans les recherches théoriques sur la variance des réponses (lire Prasad et Rao 1990 pour un exemple récent; pour les travaux antérieurs, se reporter à l'analyse exhaustive qu'on retrouve à la section 11.3 de Lessler et Kalsbeek 1992).

Puisqu'il y a pondération automatique, la moyenne de l'échantillon, \bar{Y} , correspond à l'estimateur naturel de la moyenne de la population. Sa variance selon le modèle de réponses corrélées (1) est donc:

$$V(\bar{Y}) = (n_1 \sigma_i^2 + n \sigma_p^2 + n \sigma^2) / n$$

équation dans laquelle $n_1 = \sum_i n_i^2 / n$, où n_i représente le nombre de répondants interrogés par le i -ième intervieweur et $n = \sum_i n_i$, la taille de l'échantillon, et dans laquelle $n_c = \sum_p m_p^2 / n$, où m_p indique le nombre de répondants de la p -ième UPÉ. Souignons que n_i dépasse toujours la moyenne arithmétique simple des facteurs n_i et que sa valeur peut être considérablement plus élevée si n_i varie beaucoup.

Examinons maintenant la variance prévue correspondant à $V_0(\bar{Y})$ par exemple, qu'on obtiendrait si les n observations étaient indépendantes (à savoir, si on avait prélevé un échantillon au hasard dans une très vaste population d'UPÉ au moyen d'un grand nombre d'intervieweurs). De (1), on déduit que

$$(2) \quad V_0 = \sigma_{i0}^2 / n$$

où

$$\sigma_{i0}^2 = \sigma_i^2 + \sigma_p^2 + \sigma^2.$$

Le relèvement de la variance prévue attribuable à l'effet combiné de la variabilité de l'intervieweur et de la corrélation dans le groupe est calculé au moyen du ratio

$$D_0 = V(\bar{Y}) / V_0$$

$$(3) \quad = 1 + (n_1 - 1) \rho_I + (n_c - 1) \rho_C$$

où $\rho_I = \sigma_i^2 / \sigma_{i0}^2$ et $\rho_C = \sigma_p^2 / \sigma_{i0}^2$. Nous appellerons ce ratio "effet du plan d'échantillonnage", même s'il diffère légèrement de la définition courante, laquelle se rapporte à la variance réelle plutôt qu'à la variance prévue. De l'équation (3) il ressort que la variabilité de l'intervieweur peut passablement influencer sur la variance de la moyenne d'un échantillon quand le nombre moyen de cas confiés à l'intervieweur n_i est assez important et cela, même si la corrélation intra-intervieweur ρ_I demeure assez faible.

Supposons maintenant que nous nous intéressions à l'écart entre les moyennes de deux domaines plutôt qu'à une moyenne simple, par exemple aux différences liées au sexe ou à la race. L'extension la plus simple du modèle de réponses corrélées (1) pourrait donner le modèle suivant:

$$(4) \quad Y_{ipr}^{(d)} = \mu^{(d)} + a_i + b_p + e_{ipr}^{(d)}$$

pour les observations du d -ième domaine. Dans ce cas, la moyenne $\mu^{(d)}$ de chaque domaine pourrait différer, mais on suppose que les effets attribuables à l'intervieweur et au groupe sont identiques.

Soit $p_i^{(d)} = n_i^{(d)} / n^{(d)}$, où $n_i^{(d)}$ représente le nombre de répondants du domaine d interrogés par le i -ième intervieweur, et $n^{(d)}$ correspond au nombre total de répondants du domaine d . Parallèlement, soit $q_p^{(d)} = m_p^{(d)} / n^{(d)}$, où $m_p^{(d)}$ indique le nombre de répondants du domaine d pour la p -ième UPÉ. En vertu de l'équation (4), la variance prévue de $\bar{Y}^{(d)} - \bar{Y}^{(b)}$, c'est-à-dire l'écart de la moyenne de l'échantillon pour les deux domaines serait:

$$V(\bar{Y}^{(a)} - \bar{Y}^{(b)}) =$$

$$(5) \quad (n_1 \sigma_i^2 + n \sigma_p^2 + n \sigma^2) \left(\frac{1}{n^{(a)}} + \frac{1}{n^{(b)}} \right),$$

où

$$(6) \quad m_I = \sum_i (p_i^{(a)} d_i^{(b)} - p_i^{(b)} d_i^{(a)})^2 \left(\frac{1}{n^{(a)}} + \frac{1}{n^{(b)}} \right)$$

et

$$(7) \quad m_C = \sum_p (q_p^{(a)} d_p^{(b)} - q_p^{(b)} d_p^{(a)})^2 \left(\frac{1}{n^{(a)}} + \frac{1}{n^{(b)}} \right).$$

Si les observations étaient indépendantes, la variance prévue correspondante serait:

$$V_I = \sigma_{i0}^2 \left(\frac{1}{n^{(a)}} + \frac{1}{n^{(b)}} \right)$$

si bien que l'inflation attribuable à la variabilité de l'intervieweur et à la corrélation dans le groupe s'établit désormais à:

$$D_I = \text{Var}(\bar{Y}^{(a)} - \bar{Y}^{(b)}) / V_I$$

$$(8) \quad = 1 + (m_I - 1) \rho_I + (m_C - 1) \rho_C.$$

L'ampleur de l'effet dépend de la façon dont la charge de travail des intervieweurs et les UPÉ coupent les domaines. À une extrémité, lorsque chaque intervieweur interroge le même nombre de répondants dans chaque domaine (à savoir, quand $p_i^{(a)} = p_i^{(b)}$), m_I est égal à zéro et les effets de l'intervieweur s'annulent. À l'autre cependant, lorsque

La variance de l'intervieweur et ses effets sur les comparaisons de domaines

PETER DAVIS et ALASTAIR SCOTT¹

RÉSUMÉ

Le présent article étudie les effets de la variabilité de l'intervieweur sur la précision des écarts estimés entre les moyennes de domaines. La première partie est consacrée à l'élaboration d'un modèle des éléments corrélés de la variance, permettant d'identifier les facteurs qui déterminent l'ampleur des effets, aspect qui influe sur le plan d'échantillonnage et la formation de l'intervieweur. Dans la deuxième partie, on trouvera une analyse empirique des données provenant d'une vaste enquête à plusieurs degrés sur l'hygiène dentaire. On s'est servi du sexe et de la race du répondant pour créer les deux jeux de domaines servant à la comparaison. Les effets globaux dus à l'intervieweur et au groupement de l'échantillon n'ont guère d'incidence sur la variance des comparaisons entre sexes masculin et féminin, mais la variance augmente pour certains écarts entre les deux groupes ethniques retenus. De fait, les effets de l'intervieweur sont deux ou trois fois élevés pour la comparaison des groupes ethniques que pour celle des deux sexes. Ces constatations présentent une certaine utilité pour les enquêtes dans le secteur de la santé où il est courant de recourir à un petit nombre d'intervieweurs ayant subi une formation poussée.

MOTS CLÉS: Variance due à l'intervieweur; comparaisons de domaines; effet du plan d'échantillonnage.

1. INTRODUCTION

Les enquêtes qui exigent une formation spéciale et poussée des intervieweurs, comme c'est souvent le cas dans le secteur de la santé, doivent fréquemment se contenter d'un petit nombre d'intervieweurs. Évaluer l'incidence de la variabilité de l'intervieweur sur des statistiques simples comme la moyenne et les proportions a suscité passablement de recherches, et on sait pertinemment que le recours à un nombre restreint d'intervieweurs, à qui on attribue une lourde charge de travail, peut contribuer de manière assez importante à l'erreur totale. Groves (1989, ch. 8) et Lessler et Kalisbeek (1992, n° 11.3) font une très bonne synthèse de la documentation existante à ce sujet. Néanmoins, la plupart des enquêtes médicales et sociales portent essentiellement sur des aspects plus complexes, par exemple la comparaison de sous-groupes ou l'estimation des effets d'un paramètre sur l'issue d'une maladie, si bien qu'on a tendance à croire que la variabilité de l'intervieweur est beaucoup moins importante et qu'il n'y a guère de danger à utiliser un nombre relativement restreint d'intervieweurs. Après les travaux de défrichage de Kish et Frankel (1974), les effets du groupement sur l'ajustement des modèles de régression multiple ou des modèles logarithmiques-linéaires de données catégoriques ont fait l'objet de maintes recherches théoriques et empiriques. Skinner et ses collaborateurs (1989) et Rao et Thomas (1988) ont bien dépouillé la documentation pertinente. Le problème de la comparaison des moyennes des sous-groupes, plus simple du point de vue conceptuel mais tout aussi important sur le plan pratique, a également donné lieu à certaines recherches empiriques (lire Kish 1987 et Skinner 1989, par exemple), mais on s'est relativement peu intéressé au côté théorique.

Nous examinerons surtout ici les comparaisons de sous-groupes (ou domaines). Nous débuterons par la partie théorique en utilisant un modèle simple des éléments de la variance. Théoriquement, l'impact de la variabilité de l'intervieweur dépend de deux choses: la répartition de la charge de travail de l'intervieweur entre les domaines et l'interaction entre l'intervieweur et domaine. Nous appliquerons ensuite la théorie aux données issues d'une enquête assez typique dans le secteur de la santé, en établissant deux jeux de domaines selon le sexe et la race du répondant. Malheureusement, l'enquête en question n'avait pas été conçue pour estimer les effets de l'intervieweur. Plus précisément, les intervieweurs n'avaient pas été déployés de façon aléatoire, de sorte que nos résultats sont plus indicatifs que probants. Quoiqu'il en soit, ils sont assez troublants pour que le problème justifie une étude plus poussée. Les résultats venant de la comparaison des groupes ethniques, en particulier, donnent à penser que l'utilisation d'un petit nombre d'intervieweurs pourrait susciter des difficultés dans certains cas, même lorsque l'analyse porte sur des comparaisons plutôt que des moyennes ou de simples proportions.

2. THÉORIE

Pour plus de simplicité, nous débuterons avec un plan d'échantillonnage autopondéré à deux degrés. Ce dernier est assez complexe pour mettre en relief les grandes idées, mais reste assez simple pour qu'on ne s'enlise pas dans une foule de détails. Respectant la suggestion de Collins et Butcher (1982), nous nous attaquerons simultanément au problème de la variance et à celui du groupement des

¹ Peter Davis, Department of Community Health, University of Auckland, Private Bag 92019, Auckland, New Zealand; et Professor Alastair Scott, Department of Mathematics and Statistics, University of Auckland, Private Bag 92019, Auckland, New Zealand.

Ernst et Ikeda présentent un algorithme à taille réduite pour maximiser la rétention de certaines unités primaires d'échantillonnage (UPÉ) quand un nouvel échantillon (c.-à-d. présentant une nouvelle stratification et une nouvelle répartition) est tiré pour une enquête répétée. Pour commencer, les auteurs décrivent la méthode de transport élaborée par Causey, Cox et Ernst (1985). Cette dernière permet d'optimiser la rétention des UPÉ, mais le problème de transport pourrait être trop grand pour se prêter à une solution pratique. Puis, les auteurs exposent leur algorithme, qui est une approximation de la méthode précédente, mais qui présente l'avantage d'être de taille plus petite, donc, utilisable dans de nombreuses situations pratiques. Enfin, ils présentent une application de l'algorithme à la Survey of Income and Program Participation.

Shrestha et Preston évaluent la cohérence des données sur les aînés tirées du Recensement et des registres de l'état civil aux États-Unis. Pour commencer, ils décrivent les données sur lesquelles porte l'étude et leurs sources. Puis, ils présentent la méthodologie utilisée pour évaluer la qualité des données sur les aînés et expliquent comment il convient d'interpréter les résultats de l'application de cette méthodologie. Enfin, ils présentent les résultats de l'application de la méthodologie aux données recueillies de 1970 à 1990.

Dillman, Clark et Sinclair comparent l'incidence de diverses stratégies d'envoi postal et de suivi sur le taux de réponse obtenu lors du Recensement des États-Unis. La comparaison des stratégies utilise un plan factoriel et un échantillon de 50,000 unités de logement. Les auteurs analysent les résultats en s'appuyant sur des comparaisons multiples, par paire, des moyennes de traitement et sur la régression logistique.

Forster et Snow évaluent l'utilisation d'ordinateurs portatifs pour la réalisation d'enquêtes démographiques dans les pays en développement. Une étude de collecte de données a été effectuée en vue de comparer l'utilisation de questionnaires imprimés et de questionnaires informatisés dans le cadre de l'enquête sur la mortalité des adultes menée auprès des habitants de la côte kenyane. Les résultats indiquent que l'utilisation d'ordinateurs portatifs permet de réduire le temps de traitement des données, d'améliorer la qualité de ces dernières et de diminuer les coûts d'enquête à long terme.

Le rédacteur en chef

Dans ce numéro

Ce numéro de *Techniques d'enquête* contient des articles traitant de divers sujets. Dans le premier, Davis et Scott examinent l'incidence éventuelle des effets dus à l'intervieweur sur les comparaisons entre moyennes de domaines. Grâce à un modèle des éléments de la variance, ils montrent de façon théorique que l'incidence dépend de la répartition de la charge de travail de l'intervieweur entre les domaines, d'une part, et de l'interaction entre intervieweurs et domaines, d'autre part. Ils appliquent le modèle aux données d'une enquête sur la santé afin d'estimer la grandeur des effets dus à l'intervieweur aux fins de comparaisons entre sexes et entre groupes ethniques. Les auteurs ont noté que, dans certains cas, les effets dus à l'intervieweur qui sont particuliers à un domaine ont une incidence considérable sur l'exactitude des comparaisons entre domaines.

Rivest et Hurtubise examinent l'utilité de la moyenne winsorisée comme estimateur de la moyenne d'une population ayant une distribution asymétrique vers la droite. Une moyenne winsorisée est obtenue en remplaçant toutes les observations supérieures à une valeur limite donnée R par cette même valeur R , avant le calcul de la moyenne. Les auteurs suggèrent un algorithme simple pour le calcul de R qui minimise l'erreur quadratique de l'estimateur. Les auteurs appliquent cette méthode à plusieurs tailles d'échantillon et à divers plans de sondage incluant l'échantillonnage stratifié et avec probabilités proportionnelles à la taille. Ils dérivent des approximations directes de l'efficacité de la moyenne winsorisée. Ils terminent leur article par une simulation de Monte Carlo pour comparer divers estimateurs qui réduisent l'impact des valeurs extrêmes.

Kish, Frankel et Verma examinent l'importance des effets du plan de sondage (eps) sur un ensemble de statistiques apparentées. En se fondant sur 14 enquêtes menées dans six pays, les auteurs présentent une approche empirique mettant en rapport l'effet du plan de sondage de statistiques analytiques, $\text{eps}(d_i - d_j)$, aux effets du plan de sondage des statistiques séparées, $\text{eps}(d_i)$ et $\text{eps}(d_j)$, pour deux des nombreuses catégories de la même variable. L'approximation proposée doit faire l'objet de vérifications constantes. Pourtant, elle semble s'appliquer largement aux données étudiées et est nettement préférable aux hypothèses énoncées jusqu'à maintenant sur l'eps ($d_i - d_j$). Dupont discute de l'estimation d'un total à partir d'un échantillon à deux phases et en présence d'information auxiliaire. Dans un premier temps, on présente trois estimateurs par régression chacun employant l'information auxiliaire d'une façon différente. Ensuite, quatre estimateurs par calage sont proposés chacun correspondant à une stratégie particulière d'utilisation de l'information auxiliaire. Par la suite, Dupont montre que les stratégies de calage peuvent être associées à une modélisation par régression. Dans cet article, on discute aussi de l'estimation de la variance pour les sept estimateurs présents, du choix de l'estimateur en présence de non-réponse ainsi que de l'utilisation a priori ou a posteriori de l'information auxiliaire.

Spisak discute du recours au contrôle statistique du processus pour assurer la qualité d'une base de sondage construite selon un processus en continu et utilisée pour une enquête qui se répète périodiquement. Les tailles de la base de sondage constituent une série chronologique pour laquelle il convient de déterminer le modèle qui permettra d'estimer la variance du processus nécessaire pour l'établissement des graphiques de contrôle. L'auteur se sert des données du United States Unemployment Insurance Benefits Quality Control Program pour illustrer la méthode.

Bindler et Kovacevic montrent comment on peut se servir de la méthode des équations d'estimation pour établir des procédures appropriées d'estimation de la variance dans le cas de données tirées d'une enquête dont le plan d'échantillonnage est complexe. La méthode est surtout utile lorsque la grandeur à estimer correspond à une fonction non linéaire complexe des valeurs de la population observée, comme dans le cas de bon nombre de mesures courantes de l'inégalité du revenu. Les auteurs mettent au point les détails de la méthode proposée pour plusieurs statistiques complexes de la répartition du revenu, y compris le coefficient de Gini, l'ordonnée de la courbe de Lorenz, la part du quantile et la mesure du faible revenu. Ils donnent un exemple numérique fondé sur des données de l'Enquête sur les finances des consommateurs au Canada.

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada
Volume 21, numéro 2, décembre 1995

TABLE DES MATIÈRES

Dans ce numéro	109
P. DAVIS et A. SCOTT	111
La variance de l'intervieweur et ses effets sur les comparaisons de domaines	111
L.-P. RIVEST et D. HURTUBISE	119
Moyenne winsorisée de Searls pour populations asymétriques	119
L. KISH, M.R. FRANKEL, V. VERMA et N. KAÇIROTI	131
Effets du plan de sondage sur les $(P_i - P_j)$ corrélés	131
F. DUPONT	141
Redressements alternatifs en présence de plusieurs niveaux d'information auxiliaire ...	141
D.A. BINDER et M.S. KOVACEVIC	151
Estimation de l'inégalité du revenu d'après les données d'enquête: application de la méthode des équations d'estimation	151
L.R. ERNST et M.M. IKEDA	161
Un algorithme de transport à taille réduite pour maximiser le chevauchement des enquêtes	161
D.A. DILLMAN, J.R. CLARK et M.D. SINCLAIR	173
Incidence des lettres de préavis, enveloppes-réponse affranchies et cartes de rappel sur les taux de réponse par la poste lors du recensement	173
L.B. SHRESTHA et S.H. PRESTON	181
Concordance des données censitaires et des registres de l'état civil concernant les aînés aux États-Unis: 1970-1990	181
D. FORSTER et R.W. SNOW	193
Évaluation de l'utilisation des ordinateurs de poche pour la réalisation d'enquêtes démographiques dans les pays en développement	193
A.W. SPISAK	201
Contrôle statistique du processus relatif aux bases de sondage	201
Remerciements	209

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

Techniques d'enquête est répertoriée dans The Survey Statistician et Statistical Theory and Methods Abstracts. On peut en trouver les références dans Current Index to Statistics, et Journal Contents in Qualitative Methods.

COMITÉ DE DIRECTION

Président

G.J. Brackstone

Membres

D. Binder
G.J.C. Hole
F. Mayda (Directeur de la Production)
M.P. Singh
C. Patrick

COMITÉ DE RÉDACTION

Rédacteur en chef

M.P. Singh, *Statistique Canada*

Rédacteurs associés

D.R. Bellhouse, *University of Western Ontario*

D. Binder, *Statistique Canada*

J.-C. Deville, *INSEE*

J.D. Drew, *Statistique Canada*

J.-J. Droesbeke, *Université Libre de Bruxelles*

W.A. Fuller, *Iowa State University*

M. Gonzalez, *U.S. Office of Management and Budget*

R.M. Groves, *University of Maryland*

M.A. Hidiroglou, *Statistique Canada*

D. Holt, *Central Statistical Office, U.K.*

G. Kalton, *Westat, Inc.*

A. Mason, *East-West Center*

D. Pfeffermann, *Hebrew University*

Rédacteurs adjoints

J. Denis, M. Latouche, H. Mantel et D. Stukel, *Statistique Canada*

POLITIQUE DE RÉDACTION

Techniques d'enquête publie des articles sur les divers aspects des méthodes statistiques qui intéressent un organisme statistique comme, par exemple, les problèmes de conception découlant de contraintes d'ordre pratique, l'utilisation de différentes sources de données et de méthodes de collecte, les erreurs dans les enquêtes, l'évaluation des enquêtes, la recherche sur les méthodes d'enquête, l'analyse des séries chronologiques, la désaisonnalisation, les études démographiques, l'intégration de données statistiques, les méthodes d'estimation et d'analyse de données et le développement de systèmes généralisés. Une importance particulière est accordée à l'élaboration et à l'évaluation de méthodes qui ont été utilisées pour la collecte de données ou appliquées à des données réelles. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de Statistique Canada.

Présentation de textes pour la revue

Techniques d'enquête est publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à faire parvenir le texte rédigé en anglais ou en français au rédacteur en chef, M. M.P. Singh, Division des méthodes d'enquêtes-ménages, Statistique Canada, Tunney's Pasture, Ottawa (Ontario), Canada K1A 0T6. Prière d'envoyer quatre exemplaires dactylographiés selon les directives présentées dans la revue. Ces exemplaires ne seront pas retournés à l'auteur.

Abonnement

Le prix de Techniques d'enquête (n° 12-001 au catalogue) est de 45 \$ par année au Canada, 50 \$ (É.-U.) aux États-Unis, et de 55 \$ (É.-U.) par année à l'étranger. Prière de faire parvenir votre demande d'abonnement à Section des ventes des publications, Statistique Canada, Ottawa (Ontario), Canada K1A 0T6. Un prix réduit est offert aux membres de l'American Statistical Association, l'Association Internationale de Statisticiens d'Enquête et la Société Statistique du Canada.



Ottawa

ISSN 0714-0045

N° 12-001 au catalogue

Autres pays : 55 \$ US

États-Unis : 50 \$ US

Prix : Canada : 45 \$

Décembre 1995

Tous droits réservés. Il est interdit de reproduire ou de transmettre le contenu de la présente publication, sous quelque forme ou par quelque moyen que ce soit, enregistrément sur support magnétique, reproduction électronique, mécanique, photographique, ou autre, ou de l'emmagasiner dans un système de recouvrement, sans l'autorisation écrite préalable des Services de concession des droits de licence, Division du marketing, Statistique Canada, Ottawa, Ontario, Canada K1A 0T6.

© Ministre de l'Industrie, 1995

Publication autorisée par le ministre
responsable de Statistique Canada

DÉCEMBRE 1995 • VOLUME 21 • NUMÉRO 2

UNE REVUE ÉDITÉE PAR STATISTIQUE CANADA

TECHNIQUES D'ENQUÊTE





NUMÉRO 2

•

VOLUME 21

•

DÉCEMBRE 1995

UNE REVUE
ÉDITÉE
PAR STATISTIQUE CANADA

Catégorie 12-001

TECHNIQUES D'ENQUÊTE



